ПРИКЛАДНАЯ СТАТИСТИКА

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ
СТРОИТЕЛЬНЫЙ УНИВЕРСИТЕТ

Кафедра прикладной математики

ПРИКЛАДНАЯ СТАТИСТИКА

Методические указания
к практическим занятиям для обучающихся бакалавриата
по всем УГСН 01.00.00, реализуемым НИУ МГСУ

Составители:
Т.Н. Бобылева, Л.В. Кирьянова, Т.А. Мацеевич

Москва
Издательство МИСИ – МГСУ
2020

УДК 519.2
ББК 22.17
П75

Рецензент — кандидат физико-математических наук, доцент *Н.М. Чиганова*,
доцент кафедры прикладной математики НИУ МГСУ

П75 **Прикладная статистика** [Электронный ресурс] : методические указания к практическим занятиям для обучающихся бакалавриата по всем УГСН 01.00.00, реализуемым НИУ МГСУ / сост.: Т.Н. Бобылева, Л.В. Кирьянова, Т.А. Мацеевич ; Министерство науки и высшего образования Российской Федерации, Национальный исследовательский Московский государственный строительный университет, кафедра прикладной математики. — Электрон. дан. и прогр. (0,5 Мб). — Москва : Издательство МИСИ – МГСУ, 2020. — Режим доступа: http://lib.mgsu.ru/. — Загл. с титул. экрана.

Дан материал по следующим разделам прикладной статистики: модели прикладной статистики, корреляционно-регрессионный анализ, статистика случайных процессов.

Для обучающихся бакалавриата по всем УГСН 01.00.00, реализуемым НИУ МГСУ.

Учебное электронное издание

Редактор, корректор *Е.В. Антошина*
Компьютерная верстка *О.В. Суховой*
Дизайн первого титульного экрана *Д.Л. Разумного*

Для создания электронного издания использовано:
Microsoft Word 2010, ПО Adobe Acrobat

Подписано к использованию 12.03.2020. Объем данных 0,5 Мб.

Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Национальный исследовательский
Московский государственный строительный университет»
129337, Москва, Ярославское ш., 26

Издательство МИСИ – МГСУ
Тел.: (495) 287-49-14, вн. 13-71, (499) 188-29-75, (499) 183-97-95
E-mail: ric@mgsu.ru, rio@mgsu.ru

Оглавление

МОДЕЛИ ПРИКЛАДНОЙ СТАТИСТИКИ

Дисперсионный анализ определяется как статистический метод, предназначенный для выявления влияния отдельных факторов на результат эксперимента. Модели дисперсионного анализа в зависимости от числа факторов, влияние которых изучается, подразделяются на однофакторные, двухфакторные и т.д. Например, директора фирмы интересует зависимость выполнения работ за смену от работающей на стройке бригады. Всего на стройке работают r бригад. Объём выполненных работ является результативным признаком X, работающая бригада — это фактор, который влияет на результативный признак, а уровень фактора — это конкретная бригада. Число уровней фактора r совпадает с числом работающих бригад. Здесь рассматривается однофакторная детерминированная модель дисперсионного анализа.

Суть метода **дисперсионного анализа** состоит в разложении общей вариации изучаемого показателя на части, соответствующие раздельному и совместному влиянию факторов, и статистическом изучении этих частей с целью выяснения приемлемости гипотез о существовании влияний факторов.

Пример 1. *Задача о стройке*

В таблице 1 приведены данные по объемам работ, выполненных на стройке за смену для четырех бригад.

Таблица 1

Номер бригады	Объем выполненной работы			
1	140	144	142	145
2	150	149	152	152
3	148	149	146	147
4	150	155	154	152

Проверим, выполняются ли для этих данных условия проведения дисперсионного анализа. Будем считать, что результаты выработок не зависят друг от друга и имеют нормальное распределение. Проверим по критерию Бартлетта гипотезу о равенстве групповых дисперсий. В этой задаче $r = 4$; $n_i = 4$, $i = 1; 2; 3; 4$. Объем выборки n = 16.

Расчеты проведем в таблице 2.

Таблица 2

Номер бригады (i)	Объем выполненной работы $x_{(i)j}$				Групповые средние $\overline{x}_{(i)}$	Групповые дисперсии $\tilde{S}^2_{(i)}$
1	140	144	142	145	142,75	4,92
2	150	149	152	150	150,25	1,58
3	148	149	146	147	147,50	1,67
4	150	155	154	152	152,75	4,92

Общая оценка дисперсии:

$$\tilde{S}^2 = \frac{1}{16-4}\sum_{i=1}^{4} 3\tilde{S}^2_{(i)} = 3,27\,.$$

Статистика Бартлетта:

$$t = \frac{\sum\limits_{i=1}^{4} 3\ln\dfrac{3,27}{\tilde{S}^2_{(i)}}}{1+\dfrac{1}{9}\left[\left(\sum\limits_{i=1}^{4}\dfrac{1}{3}\right)-\dfrac{1}{16-4}\right]} = 1,538\,.$$

Выберем уровень значимости α = 0,05 и по таблице распределения χ^2 с тремя степенями свободы найдем критическое значение:

$$t_{кр} = \chi^2_{0,95,\,3} = 7,82.$$

Значение статистики меньше критического, следовательно, можно принять гипотезу о равенстве групповых дисперсий.

Продолжим решение задачи о стройке. Проверим гипотезу дисперсионного анализа о равенстве средних с помощью таблицы 3.

Таблица 3

Номер бригады (i)	Групповые средние $\overline{x}_{(i)}$	Групповые дисперсии $\tilde{S}^2_{(i)}$
1	142,75	4,92
2	150,25	1,58
3	147,50	1,67
4	152,75	4,92

Общее выборочное среднее: $\overline{x} = \dfrac{1}{16}\sum_{i=1}^{4}\sum_{j=1}^{4} x_{(i)j} = 148,31.$

Межгрупповая дисперсия $S^2_M = \dfrac{1}{16}\sum_{i=1}^{4} 4\left(\overline{x}_{(i)} - 148,31\right)^2 = 13,762.$

Внутригрупповая дисперсия: $S^2_\Gamma = \dfrac{1}{16}\sum_{i=1}^{r} 3\cdot\tilde{S}^2_{(i)} = 2,454.$

Дисперсионное отношение: $F = \dfrac{S^2_M/(r-1)}{S^2_\Gamma/(n-r)} = \dfrac{12\cdot 13,762}{3\cdot 2,454} = 22,43.$

По таблицам распределения Фишера — Снедекора для α = 0,05 степеней свободы $k_1 = 3$, $k_2 = 12$ найдем критическое значение $F_{кр} = 3,49$.

Так как дисперсионное отношение больше критического, то гипотезу H_0 отклоняем и считаем, что объем ежедневной выработки зависит от работающей бригады.

Продолжим решение задачи о стройке. Ранее было получено, что объем ежедневной выработки зависит от работающей бригады. Оценим степень этой зависимости с помощью коэффициента детерминации. Было получено, что $S^2_M = 13,762$; $S^2_\Gamma = 2,454$. Следовательно, коэффициент детерминации $\overline{d} = \dfrac{S^2_M}{S^2_M + S^2_\Gamma} = \dfrac{13,762}{13,762 + 2,454} = 0,849$. Это означает, что 84,9 % общей вариации ежедневного объема выработки связано с работающей сменой.

Оценка математического ожидания (средней выработки на стройке) $m = M(X)$ равна $\tilde{m} = \overline{x} = 148,31$.

Оценки параметров $a_{(i)} = M(X\,|\,A_i) - M(X)$ соответственно равны:

$$\tilde{a}_{(1)} = \overline{x}_{(1)} - \overline{x} = 142,75 - 148,31 = -5,56;$$

$$\tilde{a}_{(2)} = \overline{x}_{(2)} - \overline{x} = 150,15 - 148,31 = 1,96;$$

$$\tilde{a}_{(i)} = \overline{x}_{(i)} - \overline{x} = 147,5 - 148,31 = -0,81;$$

$$\tilde{a}_{(i)} = \overline{x}_{(i)} - \overline{x} = 152,75 - 148,31 = 4,44.$$

Пример 2. *Задача об эффективности рекламы*

Исследователь хочет выяснить, отличаются ли три способа рекламирования товара по влиянию на объем его продажи. Для этого в каждом из трех однотипных городов (в них использовались различные способы рекламы) были собраны сведения об объемах продажи товара в пяти случайно отобранных магазинах.

Можно ли на 5% уровне значимости считать, что способ рекламы влияет на объем продаж?

Для того чтобы ответить на этот вопрос, воспользуемся моделью однофакторного дисперсионного анализа.

Здесь фактором является способ рекламы. Зафиксированы $r = 3$ его уровня и требуется выяснить, различаются ли по своему влиянию эти уровни.

Допустим, что независимость наблюдений гарантируется организацией эксперимента, а объем продаж имеет нормальное распределение.

Используя критерий Бартлетта, убедимся, что результаты испытаний позволяют принять гипотезу о равенстве групповых дисперсий.

Рассчитаем в таблице 4 выборочные характеристики:

$$\overline{x}_{(i)} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{(i)j}\,,\ \tilde{S}^2_{(i)} = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{(i)\,j} - \overline{x}_{(i)})^2\,,$$

здесь $n_i = 5$, $n = 15$.

Таблица 4

Способ рекламы (i)	Объем продаж $x_{(i)\,j}$					Групповые средние $\overline{x}_{(i)}$	Групповые дисперсии $\tilde{S}^2_{(i)}$
1-ый	2	2	6	3	3	3,2	4,66
2-ой	8	6	4	4	9	6,2	5,2
3-ий	7	7	6	3	2	5,0	5,5

Вычислим общую оценку дисперсии:

$$\tilde{S}^2 = \frac{1}{n-r}\sum_{i=1}^{r}(n_i - 1)\tilde{S}^2_{(i)} = \frac{1}{15-3}\sum_{i=1}^{3} 4\tilde{S}^2_{(i)} = 5{,}12\,.$$

Статистика Бартлетта:

$$t = \frac{\sum_{i=1}^{r}(n_i - 1)\ln\dfrac{\tilde{S}^2}{\tilde{S}^2_{(i)}}}{1 + \dfrac{1}{3(r-1)}\left[\left(\sum_{i=1}^{r}\dfrac{1}{n_i - 1}\right) - \dfrac{1}{n-r}\right]} = \frac{\sum_{i=1}^{3} 4\ln\dfrac{5{,}12}{\tilde{S}^2_{(i)}}}{1 + \dfrac{1}{6}\left[\left(\sum_{i=1}^{4}\dfrac{1}{4}\right) - \dfrac{1}{15-3}\right]} = 0{,}03\,.$$

Выберем уровень значимости $\alpha = 0{,}05$ и по таблице распределения χ^2 с тремя степенями свободы найдем критическое значение: $t_{кр} = \chi^2_{0{,}95,\,3} = 7{,}82$.

Значение статистики меньше критического, следовательно, можно принять гипотезу о равенстве групповых дисперсий.

Теперь проверим гипотезу дисперсионного анализа о равенстве средних.

Общее выборочное среднее: $\overline{x} = \dfrac{1}{15}\sum_{i=1}^{3}\sum_{j=1}^{5} x_{(i)\,j} = 4{,}8\,.$

Межгрупповая дисперсия: $S^2_M = \dfrac{1}{n}\sum_{i=1}^{r} n_i\left(\overline{x}_{(i)} - \overline{x}\right)^2 = \dfrac{1}{15}\sum_{i=1}^{3} 5\left(\overline{x}_{(i)} - 4{,}8\right)^2 = 1{,}52\,.$

Внутригрупповая дисперсия: $S^2_\Gamma = \dfrac{1}{n}\sum_{i=1}^{r}(n_i - 1)\cdot\tilde{S}^2_{(i)} = \dfrac{1}{15}\sum_{i=1}^{3} 4\cdot\tilde{S}^2_{(i)} = 4{,}096\,.$

Дисперсионное отношение: $F = \dfrac{S^2_M/(r-1)}{S^2_\Gamma/(n-r)} = \dfrac{12\cdot 1{,}52}{2\cdot 4{,}096} = 2{,}23\,.$

По таблицам распределения Фишера — Снедекора для $\alpha = 0{,}05$ степеней свободы $k_1 = r - 1 = 2$, $k_2 = n - r = 12$ найдем критическое значение $F_{кр} = 3{,}89$.

Так как дисперсионное отношение меньше критического, то гипотезу о равенстве средних принимаем и считаем, что объем продаж не зависит от способа рекламы.

Пример 3. *Задача об экспертных оценках*

Каждый из 10 экспертов оценивал 20 случайно отобраных студенческих работ по определенной шкале. Общая выборочная дисперсия $S^2 = 9{,}16$. Межгрупповая дисперсия (дисперсия, обусловленная различиями экспертов) $S_M^2 = 0{,}47$. Кто (или что) в среднем отличается больше: эксперты или студенческие работы, которые оценивает один и тот же эксперт?

Условия задачи не дают данных для проверки гипотезы о равенстве групповых дисперсий (по критерию Бартлетта), предполагая, что все условия применения дисперсионного анализа выполнены.

В решаемой задаче результативный признак X — это оценка эксперта, выставленная за определенную работу. Фактор, влияющий на оценку, — сам эксперт, поэтому число уровней фактора $r = 10$. Группа данных, связанная с определенным уровнем фактора — это оценки, выставленные экспертом, поэтому объем выборки по одной группе $n_i = 20$. Общий объем данных $n = 20 \cdot 10 = 200$ (общее число работ, проверенных всеми экспертами).

Проверим гипотезу о равенстве групповых средних, для этого определим внутригрупповую дисперсию:

$$S_Г^2 = S^2 - S_M^2 = 9{,}16 - 0{,}47 = 8{,}69.$$

Дисперсионное отношение:

$$F = \frac{S_M^2 / (r-1)}{S_Г^2 / (n-r)} = \frac{(200-10) \cdot 0{,}47}{(10-1) \cdot 8{,}69} = 1{,}142.$$

По таблицам распределения Фишера — Снедекора для $\alpha = 0{,}05$ степеней свободы $k_1 = r - 1 = 9$, $k_2 = n - r = 190$ найдем критическое значение Fкр = 1,83.

Так как дисперсионное отношение меньше критического, то гипотезу о равенстве средних принимаем и считаем, что выставленная оценка не зависит от личности эксперта. Таким образом, в среднем студенческие работы, которые оценивает один и тот же эксперт, отличаются больше, чем оценки различных экспертов по одинаковым работам.

КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

Основной задачей корреляционного анализа является выявление статистической зависимости между случайными переменными путём оценок различных коэффициентов корреляции.

В теории вероятностей и математической статистике изучается, как правило, **стохастическая** зависимость между случайными величинами, когда одному и тому же значению $\boldsymbol{X}$ могут соответствовать, в зависимости от случая, различные значения величины $\boldsymbol{Y}$. При стохастической зависимости величины не связаны функционально, но как случайные величины связаны совместным распределением вероятности. Наличие стохастической зависимости объясняется тем, что на результирующую переменную $\boldsymbol{Y}$ действует на только контролируемый фактор $\boldsymbol{X}$, но и множество других неконтролируемых случайных факторов.

Корреляционной зависимостью между переменными называется функциональная зависимость между значениями одной из них и условным математическим ожиданием другой.

Пример 1. Исследуем корреляционную зависимость между суточной выработкой продукции (Y тонн) и величиной основных производственных фондов (X млн. руб.): найдем коэффициент корреляции, проверим его значимость и построим для него доверительный интервал. Данные уже сгруппированы, в качестве значений x_i и y_j приведены середины интервалов.

Таблица 5

	$y_1 = 9$	$y_2 = 13$	$y_3 = 17$	$y_4 = 21$	$y_5 = 25$	Всего ($n_{i\bullet}$)
$x_1 = 22.5$	2	1	–	–	–	3
$x_2 = 27.5$	3	6	4	–	–	13
$x_3 = 32.5$	–	3	11	7	–	21
$x_4 = 37.5$	–	1	2	6	2	11
$x_5 = 42.5$	–	–	–	1	1	2
Всего ($n_{\bullet j}$)	5	11	17	14	3	50 (= $\boldsymbol{n}$)

Используя данные, приведённые в таблице 5, находим:

1) выборочные средние $\overline{x} = 32{,}1$ (млн. руб.), $\overline{y} = 16{,}92$ (тонн);
2) выборочные дисперсии $S_X^2 = 21{,}84$; $S_Y^2 = 18{,}23$;
3) эмпирический коэффициент корреляции $\hat{r} = \dfrac{14{,}768}{\sqrt{21{,}84 \cdot 18{,}23}} = 0{,}740.$

Проверим на уровне $\alpha = 0{,}05$ значимость полученного коэффициента корреляции.

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

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:

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

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

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

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

Пример 2. Продолжим решение задачи о зависимости между суточной выработкой продукции (Y тонн) и величиной основных производственных фондов (X млн. руб.). Найдем корреляционное отношение между исследуемыми величинами и проверим, можно ли считать связь между величинами линейной.

Таблица 6

	$y_1 = 9$	$y_2 = 13$	$y_3 = 17$	$y_4 = 21$	$y_5 = 25$	Всего ($n_{i\bullet}$)	Групповые средние ($\overline{y}_i$)
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 (=$\boldsymbol{n}$)	
Групповые средние ($\overline{x}_j$)	25.5	29.3	31.9	35.4	39.2		

Вычислим эмпирическое корреляционное отношение:

$$\hat{\eta}_{Y|X} = \frac{1}{S_Y}\sqrt{\frac{1}{50}\sum_{i=1}^{5}\left(\overline{y}_i - \overline{y}\right)^2 \cdot n_{i\bullet}} = \sqrt{\frac{10,36}{18,23}} = 0,754 .$$

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 $l = 5$, вычислим значение статистики:

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

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

Ранговая корреляция

В том случае, когда изучаются не количественные признаки**,** а качественные, такие меры зависимости как коэффициент корреляции и корреляционное отношение не годятся. Однако часто удаётся как-то упорядочить изучаемые объекты в отношении некоторого признака, то есть приписать им порядковые номера — **ранги**.

Как правило, нумерация (присвоение ранга) идет от **1** до $\boldsymbol{n}$ по возрастанию значений признака. Если встречается несколько одинаковых значений $\boldsymbol{x_i}$ или $\boldsymbol{y_j}$, то каждому из них присваивается ранг, равный частному от деления суммы рангов, приходящихся на эти значения, на число равных значений.

Ранги признаков $\boldsymbol{X}$ и $\boldsymbol{Y}$ обозначают N_x и N_y . В качестве выборочной характеристики связи можно воспользоваться **ранговым коэффициентом корреляции Спирмена**: $\rho = 1 - \dfrac{6\cdot\sum_{i=1}^{n} d_i^2}{n\cdot(n^2-1)}$, где $d_i = N_x - N_y$ — разность рангов по обоим признакам для каждого объекта. Ранговый коэффициент корреляции Спирмена может принимать значения от –1 до +1. 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 $\boldsymbol{n} > 10$ для вычисления критических значений выборочных коэффициентов пользуются тем, что эти величины распределены приближенно нормально.

Пример 3. Экспертами оценивались качества разных вин. Согласуется ли оценка качества вина с его ценой? Суммарные оценки получены следующие (Таблица 7):

Таблица 7

Марка вина	1	2	3	4	5	6	7	8	9	10
Оценка в баллах (***X***)	11	14	17	15	13	13	18	10	19	25
Цена в усл. ед. (***Y***)	1,57	1,60	2,00	2,10	1,70	1,85	1,80	1,15	2,30	2,40

Найдем ранги признаков и разность рангов. Вычисления проведем с помощью таблицы 8. В первых двух столбцах перепишем исходные данные, третий столбец — ранг величины ***X*** (оценки в баллах), четвертый — ранги ***Y*** (цена). Наименьший ранг равен единице, он соответствует самой маленькой оценке (для ***X***) и самой маленькой цене (для ***Y***).

Таблица 8

X	***Y***	N_x	N_y	$d = N_x - N_y$	d^2
14	1,60	5	3	2	4
17	2,00	7	7	0	0
15	2,10	6	8	-2	4
13	1,70	3,5	4	0,5	0,25
13	1,85	3,5	6	2,5	6,25
18	1,80	8	5	3	9
10	1,15	1	1	0	0
19	2,30	9	9	0	0
25	2,40	10	10	0	0

$$\sum d^2 = 23,5.$$

Далее найдём значение коэффициента Спирмена:

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot 23,5}{10 \cdot (100 - 1)} \approx 0,8576 \quad \rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot 23,5}{10 \cdot (100 - 1)} \approx 0,8576.$$

По таблице значений коэффициента корреляции рангов Спирмена определяем, что при объеме выборки 10 единиц (***n*** = 10) и уровне значимости 5% (α = 0,05) критическая величина для рангового коэффициента корреляции составляет $\pm 0,6364$. Найденное значение коэффициента ρ больше критического. Таким образом, можно сделать вывод о весьма большой тесноте прямой зависимости между качеством вина и его стоимостью.

Пример 4. По территориям России имеются данные, представленные в таблице 9 выборочных парных коэффициентов корреляции.

Таблица 9

Выборочные коэффициенты корреляции	Среднедневная заработная плата, руб., X	Средний возраст безработного, лет, Y	Среднедушевой доход, руб., Z
Среднедневная заработная плата, руб., X	1	–0,1160	0,8405
Средний возраст безработного, лет, Y	–0,1160	1	–0,2101
Среднедушевой доход, руб., Z	0,8405	–0,2101	1

Рассчитаем множественный и частные коэффициенты корреляции и выборочный множественный коэффициент детерминации: $R_{Z,X} = 0,8404$, $R_{Z,Y} = -0,2092$, $R_{X,Y} = 0,1144$, $R_Z = 0,8481$, $R_Z^2 = 0,7193 \approx$ $\approx 0,72$.

Выводы:

1) из-за слабой связи между ***X*** и ***Y*** коэффициенты парной и частной корреляции отличаются незначительно;
2) зависимость Z от X и Y характеризуется как тесная, в которой 72% изменения среднего душевого дохода определяются изменением учтённых факторов: средней заработной платы и среднего возраста безработного.

Парная линейная регрессия.

Пример 1. В магазине постельных принадлежностей в течение 5 дней подсчитывали число покупок простыней ***X*** и подушек ***Y*** (Таблица 10).

Таблица 10

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

Найти выборочный коэффициент корреляции и выборочное уравнение линейной регрессии.
Транспонируем и расширим таблицу для упрощения подсчетов (Таблица 11).

Таблица 11

	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:

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

Находим значение выборочного коэффициента корреляции:

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

Посчитаем выборочные коэффициенты линейной регрессии

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

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

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:

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

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

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

Отсюда $\hat{\sigma} \approx 2,1$.

Зададим уровень значимости ***p*** = 0,1. Тогда критическая граница $t_p = 2,35$ и доверительный интервал имеет вид:

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

СТАТИСТИКА СЛУЧАЙНЫХ ПРОЦЕССОВ

Стационарные случайные процессы

Случайный процесс — семейство случайных величин, индексированных некоторым параметром, чаще всего играющим роль времени или координаты.

Случайный процесс называется стационарным, если все его характеристики не зависят от времени. Случайная функция $X(t)$ называется стационарной в широком смысле, если ее математическое ожидание постоянно, а корреляционная функции зависит только от разности аргументов. Из этого предположения следует возможность наилучшей в среднем квадратичном линейной интерполяции, экстраполяции и фильтрации.

Статистический анализ случайных процессов посвящен методам обработки и использования статистических данных, относящихся к случайным процессам. Данные о случайном процессе $\xi(t)$, используемые при статистическом анализе этого процесса, обычно представляют собой сведения о значениях одной или нескольких реализаций $x(t)$ данного процесса в течение определенного промежутка времени. При решении задач прогнозирования случайных процессов возникают задачи статистической оценки неизвестных параметров процесса $\xi(t)$. Это происходит тогда, когда по данным наблюдений за траекторией $x(t)$ процесса $\xi(t)$ в течение определенного промежутка времени требуется оценить значение процесса $\xi(t)$ в фиксированный момент времени. Экономический временной ряд: $y(t) = f(t) + \xi(t)$, $t = 1, 2, ..., T$, где $y(t)$ — значение временного ряда; $f(t)$ — детерминированная составляющая — **тренд**; $\xi(t)$ — случайная составляющая; T — длина временного ряда.

Роль детерминированной составляющей часто играет результирующий показатель, представляющий собой, например, объем производства, обусловленный общей тенденцией экономического роста, научно-техническим прогрессом и затратами экономических ресурсов. Случайная же составляющая аккумулирует влияние множества не включенных в детерминированную составляющую факторов, каждый из которых отдельно оказывает незначительное влияние на результат.

Пример 1. Дан случайный процесс $X(t) = \cos(t + \varphi)$, где φ — случайная величина, равномерно распределенная на отрезке $[0; 2\pi]$. Требуется доказать, что этот случайный процесс стационарен в широком смысле.

Решение. Найдем математическое ожидание

$$m_X(t) = M[X(t)] = M[\cos(t + \varphi)] = M[\cos t \cos\varphi - \sin t \sin\varphi] =$$
$$= \cos t\, M[\cos\varphi] - \sin t\, M[\sin\varphi] = 0,$$

так как

$$M[\cos\varphi] = \frac{1}{2\pi}\int_0^{2\pi} \cos\varphi\, d\varphi = \frac{1}{2\pi}\sin\varphi\Big|_0^{2\pi} = 0,\ M[\sin\varphi] = \frac{1}{2\pi}\int_0^{2\pi} \sin\varphi\, d\varphi = -\frac{1}{2\pi}\cos\varphi\Big|_0^{2\pi} = 0,$$

где $\frac{1}{2\pi}$ — плотность вероятности случайной величины φ.

Заметим, что $\overset{o}{X}(t) = X(t) - m_X(t) = \cos(t + \varphi) - 0 = \cos(t + \varphi)$

$$K_X(t_1, t_2) = M[\overset{o}{X}(t_1)\,\overset{o}{X}(t_2)] = M[\cos(t_1 + \varphi)\cos(t_2 + \varphi)] =$$
$$= \frac{1}{2}M[\cos(t_2 + \varphi - t_1 - \varphi) + \cos(t_1 + \varphi + t_2 + \varphi)] =$$
$$= \frac{1}{2}\cos(t_2 - t_1) + \frac{1}{2}M[\cos(t_2 + t_1 + 2\varphi)] = \frac{1}{2}\cos(t_2 - t_1),$$

так как

$$M[\cos\,(t_2+t_1+2\varphi)]=\frac{1}{2\pi}\int_0^{2\pi}\cos(t_2+t_1+2\varphi)\;d\varphi=\frac{1}{4\pi}\sin(t_2+t_1+2\varphi)\,|_0^{2\pi}=$$

$$=\frac{1}{4\pi}[\sin(t_2+t_1+4\pi)-\sin(t_2+t_1)]=0.$$

Итак, $m_X(t)=0$, а $K_X(t_1,t_2)=\frac{1}{2}\cos(t_2-t_1)$, то есть зависит только от разности $t_2-t_1=\tau$. Корреляционная функция оказалась не зависящей от величины φ, которую в приложениях обычно называют фазой.

Аналитическое выравнивание временных рядов. Оценка параметров уравнения тренда

Одной из важнейших задач исследования экономического временного ряда является выявление основной тенденции изучаемого процесса, выраженной неслучайной составляющей (тренда).

Пример 2. По данным таблицы 12 найти уравнение неслучайной составляющей (тренда) для y_t, полагая тренд линейным.

Таблица 12

Год t	1	2	3	4	5	6	7	8
Спрос y_t	213	171	291	309	317	362	351	361

Решение. Предположим, что динамика ряда описывается линейной функцией $y=b_0+b_1t$. Тогда, согласно методу наименьших квадратов, для определения неизвестных параметров b_0 и b_1 используется система нормальных уравнений:

$$\begin{cases} b_0+b_1\sum\limits_{i=1}^{n}t_i=\sum\limits_{i=1}^{n}(y_t)_i \\ b_0\sum\limits_{i=1}^{n}t_i+b_1\sum\limits_{i=1}^{n}t_i^2=\sum\limits_{i=1}^{n}t_i(y_t)_i \end{cases}.$$

Учитывая, что значения переменной t образуют последовательность натуральных чисел от 1 до n, суммы t и t^2 можно вычислить по формулам $\sum\limits_{i=1}^{n}t_i=\frac{n(n+1)}{2}$, $\sum\limits_{i=1}^{n}t_i^2=\frac{n(n+1)(2n+1)}{6}$, имеем

$$\sum_{i=1}^{8}t_i=\frac{8\cdot9}{2}=36,\ \sum_{i=1}^{8}t_i^2=\frac{8\cdot9\cdot17}{6}=204,\ \sum_{i=1}^{8}(y_t)_i=2375,\ \sum_{i=1}^{8}t_i(y_t)_i=11766.$$

Система нормальных уравнений имеет вид:

$$\begin{cases} 8b_0+36b_1=2375 \\ 36b_0+204b_1=11766 \end{cases}.$$

Откуда $b_0=181{,}32$; $b_1=25{,}679$ и уравнение тренда $\hat{y}_t=181{,}32+25{,}679t$, то есть спрос ежегодно увеличивается в среднем на 25,7 единиц.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Постовалов С.Н. Математическая статистика: учебное пособие / Постовалов С.Н., Чимитова Е.В., Карманов В.С. — Новосибирск: Новосибирский государственный технический университет, 2014. — URL: http://www.iprbookshop.ru/45381.html.(дата обращения 17.04.2018). — Режим доступа: ЭБС «IPRbooks».
2. Прохоров Ю.В., Пономаренко Л.С. Лекции по теории вероятностей и математической статистике / Ю.В. Прохоров, Л.С. Пономаренко. — Москва: Издательство Юрайт, 2019. — 219 с. — ISBN 978-5-534-10807-1.
3. Мхитарян В.С. Теория вероятностей и математическая статистика / В.С. Мхитарян и др. — Москва: Академия, 2012. — 412 с. — ISBN 978-5-7695-8147-2.