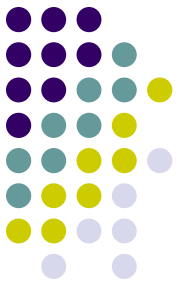




Дисциплина:

## МАТЕМАТИКА ППИ

- Лектор: Ахкамova Юлия Абдулловна
- доцент кафедры математики и методики обучения математике ЮУрГГПУ
- [akhkamovayua@cspu.ru](mailto:akhkamovayua@cspu.ru)

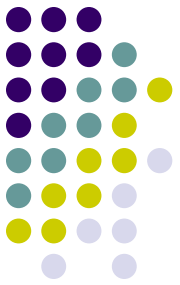


Учебный вопрос.

Корреляция. Коэффициент

корреляции. Основы

регрессионного анализа



# ПОДВОПРОСЫ

- **1.** Корреляция. Коэффициент корреляции.
- **2.** Основы регрессионного анализа

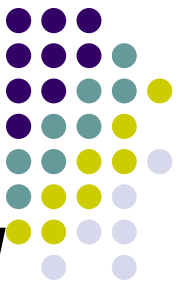
# ПОДВОПРОС



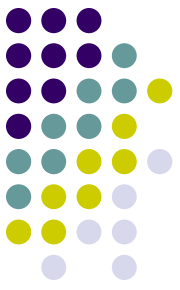
- **Корреляция. Коэффициент корреляции.**



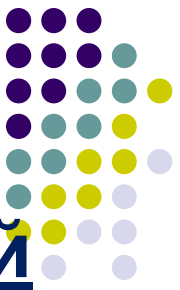
- **Применение статистических методов при обработке материалов психологических исследований дает большую возможность извлечь из экспериментальных данных полезную информацию.**
- **Одним из самых распространенных методов статистики является корреляционный анализ.**



*Термин «корреляция» впервые применил французский палеонтолог Ж. Кювье, который вывел «закон корреляции частей и органов животных» (этот закон позволяет восстанавливать по найденным частям тела облик всего животного). В статистику указанный термин ввел английский биолог и статистик Ф. Гальтон (не просто «связь» – *relation*, а «как бы связь» – *corelation*).*

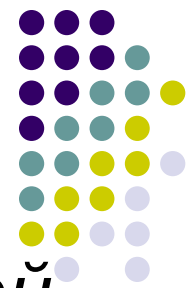


- **Корреляционный анализ** занимается степенью связи между двумя случайными величинами  $X$  и  $Y$ .
- **Основные приемы корреляционного анализа:**
  - 1.) Вычисление выборочных коэффициентов корреляции.
  - 2.) Составление корреляционной таблицы.
  - 3.) Проверка статистической гипотезы значимости связи.

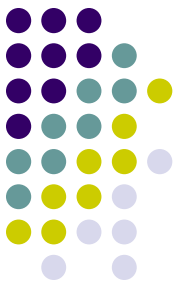


- **ОПРЕДЕЛЕНИЕ.** Корреляционная зависимость между случайными величинами  $X$  и  $Y$  называется линейной корреляцией, если обе функции регрессии являются линейными. В этом случае обе линии регрессии являются прямыми; они называется прямыми регрессии.
- Для достаточно полного описания особенностей корреляционной зависимости между величинами недостаточно определить форму этой зависимости и в случае линейной зависимости оценить ее силу по величине коэффициента регрессии.





*Например, ясно, что корреляционная зависимость возраста  $Y$  учеников средней школы от года  $X$  их обучения в школе является, как правило, более тесной, чем аналогичная зависимость возраста студентов высшего учебного заведения от года обучения, поскольку среди студентов одного и того же года обучения в вузе обычно наблюдается больший разброс в возрасте, чем у школьников одного и того же класса.*

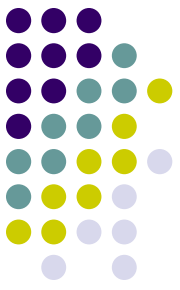


Для оценки тесноты линейных корреляционных зависимостей между величинами  $X$  и  $Y$  по результатам выборочных наблюдений **вводится понятие выборочного коэффициента линейной корреляции, определяемого формулой:**

$$r_B = \frac{\overline{XY} - \bar{X}\bar{Y}}{\sigma_X \sigma_Y} \quad (1)$$

где  $\sigma_X$  и  $\sigma_Y$  -выборочные средние квадратические отклонения величин  $X$  и  $Y$ , которые вычисляются по формулам:

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{\overline{X^2} - (\bar{X})^2}, \quad \sigma_Y = \sqrt{\sigma_Y^2} = \sqrt{\overline{Y^2} - (\bar{Y})^2}, \quad \bar{Y}^2 = \frac{1}{n} \sum_{j=1}^k n_{y_j} y_j^2 \quad (2)$$



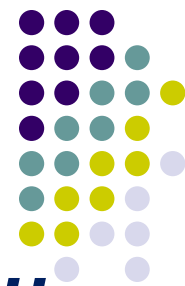
- Следует отметить, что основной смысл выборочного коэффициента линейной корреляции  $r_B$  состоит в том, что он представляет собой эмпирическую (т.е. найденную по результатам наблюдений над величинами  $X$  и  $Y$ ) оценку соответствующего генерального коэффициента линейной корреляции  $r$ :

- $r = r_B$  (3)

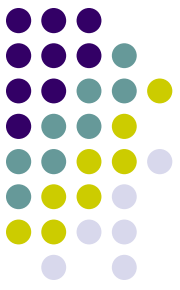
- **Выборочное уравнение линейной регрессии  $Y$  на  $X$  имеет вид:**  $Y - \bar{Y} = r_B \frac{\sigma_Y}{\sigma_X} (X - \bar{X})$  (4)

где  $r_B \frac{\sigma_Y}{\sigma_X} = b$ . То же можно сказать о выборочном уравнении линейной регрессии  $X$  на  $Y$ :

$$X - \bar{X} = r_B \frac{\sigma_X}{\sigma_Y} (Y - \bar{Y}) \quad (5)$$



- **Основные свойства выборочного коэффициента линейной корреляции:**
  1. Коэффициент корреляции двух величин, не связанных линейной корреляционной зависимостью, равен нулю.
  2. Коэффициент корреляции двух величин, связанных линейной корреляционной зависимостью, равен 1 в случае возрастающей зависимости и -1 в случае убывающей зависимости.



- **3. Абсолютная величина коэффициента корреляции двух величин, связанных линейной корреляционной зависимостью, удовлетворяет неравенству  $0 < |r| < 1$ . При этом коэффициент корреляции положителен, если корреляционная зависимость возрастающая, и отрицателен, если корреляционная зависимость убывающая.**
- **4. Чем ближе  $|r|$  к 1, тем теснее прямолинейная корреляция между величинами  $Y$ ,  $X$ .**
- **По своему характеру корреляционная связь может быть прямой и обратной, а по силе – сильной, средней, слабой. Кроме того, связь может отсутствовать или быть полной.**



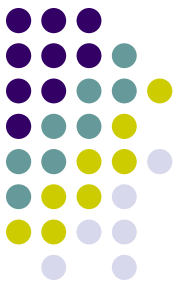
# Сила и характер связи между параметрами

Сила связи	Характер связи	
	Прямая (+)	Обратная (-)
Полная	1	-1
Сильная	От 0,7 до 1	От -0,7 до -1
Средняя	От 0,3 до 0,7	От -0,3 до -0,7
Слабая	От 0,3 до 0	От -0,3 до 0
Связь отсутствует	0	0

**Пример 1.** Изучалась зависимость между качеством  $Y$  (%) и количеством  $X$  (шт). Результаты наблюдений приведены в виде корреляционной таблицы:



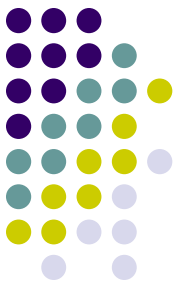
$Y \setminus X$	18	22	26	30	$n_y$
70	5				$5=5+0+0+0$
75	7	46	1		$54=7+46+1+0$
80		29	72		$101=0+29+72+0$
85			29	8	$37=0+0++29+8$
90				3	$3=0+0+0+3$
$n_x$	$12=5+7+0+0+0$	$75=0+46+29+0+0$	$102=0+1+72+29+0$	$11=0+0+0+8+3$	$200=200$



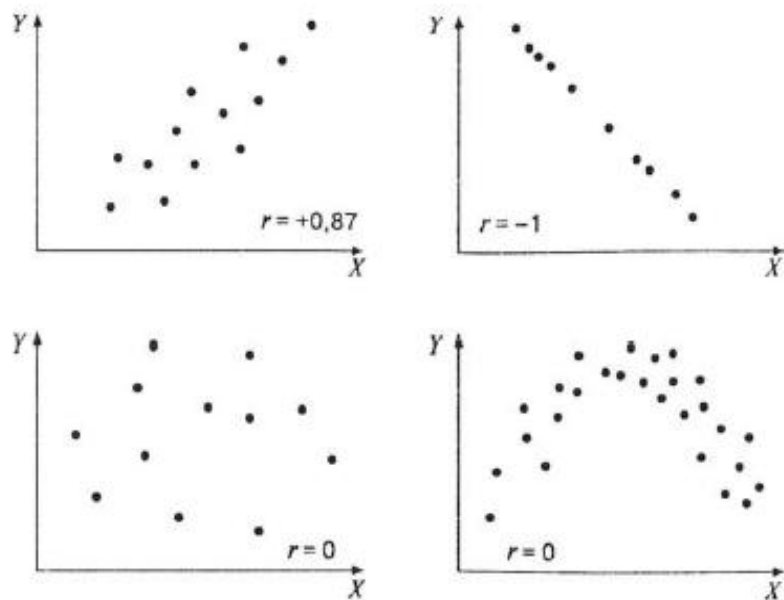
- **Требуется вычислить выборочный коэффициент линейной корреляции зависимости  $Y$  от  $X$ .**
- *Решение.*  $\bar{x}_B = 24,24$ ;  $\bar{y}_B = 79,475$ ;
- $s_x = 7,68$ ;  $s_y = 15,25$ .
- $\overline{xy}$  найдем по таблице, которую составлять будем на практическом занятии.
- $r_B = 0,7837$ ;
- $y - \bar{y}_B = r_B \frac{s_y}{s_x} (x - \bar{x}_B)$
- $y - 79,475 = 0,7837 \cdot \frac{15,25}{7,68} (x - 24,24)$
- $y - 79,475 = 0,7837 \cdot 1,98 \cdot (x - 24,24)$
- $y - 79,475 = 1,5517 \cdot (x - 24,24)$
- $y - 79,475 = 1,5517x - 1,5517 \cdot 24,24$
- $y - 79,475 = 1,5517x - 37,61$



- **Вывод: Корреляционная зависимость между величинами  $X$  и  $Y$  –**
- **прямая и сильная.**
- **По форме корреляционная связь может быть линейной или нелинейной.**
- **Для линейной корреляционной связи можно выделить два основных направления: положительное («прямая связь») и отрицательное («обратная связь»).**
- **Сила связи напрямую указывает, насколько ярко проявляется совместная изменчивость изучаемых переменных.**



В психологии функциональная взаимосвязь явлений эмпирически может быть выявлена только как вероятностная связь соответствующих признаков. Наглядное представление о характере вероятностной связи дает диаграмма рассеивания – график, оси которого соответствуют значениям двух переменных, а каждый испытуемый представляет собой точку.



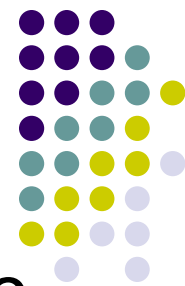
Примеры рассеивания и соответствующих коэффициентов корреляции



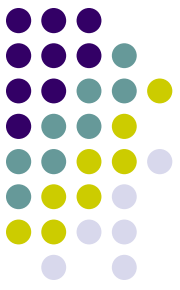
- В малых выборках для дальнейшей интерпретации корректнее отбирать сильные корреляции на основании уровня статистической значимости.
- Для исследований, которые проведены на больших выборках, лучше использовать абсолютные значения коэффициентов корреляции.
- Основная статистическая гипотеза, которая проверяется корреляционным анализом, является ненаправленной и содержит утверждение о равенстве корреляции нулю в генеральной совокупности  $H_0: r_{xy} = 0$ .



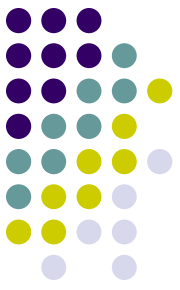
- При ее отклонении принимается альтернативная гипотеза  $H_1: r_{xy} \neq 0$  о наличии положительной или отрицательной корреляции – в зависимости от знака вычисленного коэффициента корреляции.
- На основании принятия или отклонения гипотез делаются содержательные выводы.
- Однако к интерпретации выявленных корреляционных связей следует подходить осторожно.
- С научной точки зрения, простое установление связи между двумя переменными не означает существования причинно-следственных отношений.



- *Существует множество ситуаций, в которых его применение целесообразно. Например: установление связи между интеллектом школьника и его успеваемостью;*
- *между настроением и успешностью выхода из проблемной ситуации;*
- *между уровнем дохода и темпераментом и т. п.*
- Коэффициент Пирсона находит широкое применение в психологии и педагогике.



- При вычислениях на компьютере статистическая программа (SPSS, Statistica) сопровождает вычисленный коэффициент корреляции более точным значением  $p$ -уровня.
- Для статистического решения о принятии или отклонении  $H_0$  обычно устанавливают  $\alpha = 0,05$ , а для большого объема наблюдений (100 и более)  $\alpha = 0,01$ .



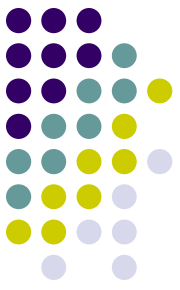
- Если  $p \leq \alpha$ ,  $H_0$  отклоняется и делается содержательный вывод, что обнаружена статистически достоверная (значимая) связь между изучаемыми переменными (положительная или отрицательная – в зависимости от знака корреляции). Когда  $p > \alpha$ ,  $H_0$  не отклоняется, содержательный вывод ограничен констатацией, что связь (статистически достоверная) не обнаружена.
- Если связь не обнаружена, но есть основания полагать, что связь на самом деле есть, следует проверить возможные причины недоверности связи.

# УЧЕБНЫЙ ВОПРОС



- **Основы регрессионного анализа**



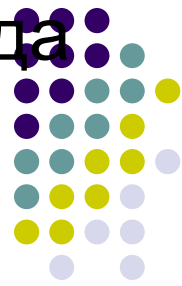


- **Задачи, решаемые методами регрессии и корреляции, непосредственно связаны между собой.** В то время, как в корреляционном анализе оценивается интенсивность, теснота связи, в регрессионном анализе исследуется ее форма. Иногда регрессию рассматривают как частный случай корреляции, считая тем самым корреляцию более широким понятием.
- Корреляция в широком смысле слова означает связь, соотношение между объективно существующими явлениями и процессами. **Не каждую корреляцию можно отождествлять с причинной связью.** При изучении совместного изменения явлений может быть установлена так называемая ложная корреляция.

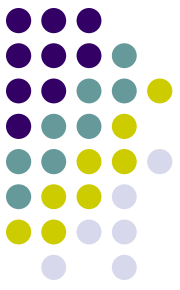


- Под ложной корреляцией понимается **чисто количественная сопряженность в вариации изучаемых явлений, не имеющая логического объяснения по содержанию.**
- Для эффективного изучения связи необходимо использовать совокупности единиц достаточно большого объема и однородные в отношении тех признаков, связь которых изучается.

Прямолинейная зависимость имеет место, когда с возрастанием (или убыванием) значений признака-фактора значения результативного признака увеличиваются (или уменьшаются) более ли менее равномерно. Линейное уравнение парной регрессии:  $\hat{y}_x = a + bx$



- где  $\hat{y}_x$  среднее значение результативного признака при определенном значении факторного признака  $x$ ;
- $a$  – свободный член уравнения регрессии;
- $b$  – коэффициент регрессии, который показывает, на сколько единиц в среднем изменится результативный признак  $y$  при изменении факторного признака  $x$  на одну единицу его измерения.

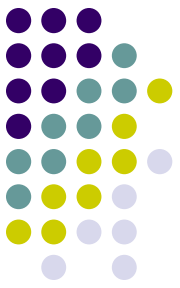


- Криволинейная форма связи может выражаться различными видами функций, из которых наиболее часто используются парабола второго порядка, гипербола, показательная, степенная.
- С целью проверки качества модели связи используются математические критерии адекватности.
- Оценки неизвестных параметров уравнения регрессии находят обычно методом наименьших квадратов (МНК):

$$f(a, b) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n [y_i - (a + bx_i)]^2 \rightarrow \min$$

Система нормальных уравнений МНК для прямой:

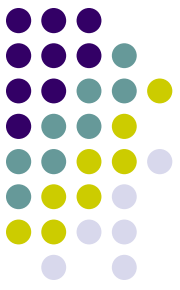
$$\begin{cases} na + b \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i x_i \end{cases}$$



Отсюда:  $a = \frac{\Delta_a}{\Delta}; b = \frac{\Delta_b}{\Delta}$

где  $\Delta$  – определитель системы;  $\Delta_a$  – частный определитель, получаемый путем замены коэффициентов при  $a$  членами правой части системы уравнений;  $\Delta_b$  – частный определитель, получаемый путем замены коэффициентов при  $b$  членами правой части системы уравнений.

$$\Delta = n \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n y_i$$



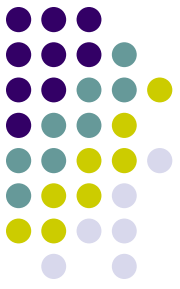
Тогда

$$a = \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i x_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2};$$

$$b = \frac{n \sum_{i=1}^n y_i x_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2}.$$

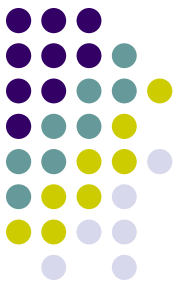
Параметры  $a$  и  $b$  могут быть выражены следующим образом:  $a = \bar{y} - b\bar{x}$

$$b = \frac{\overline{yx} - \bar{y} \cdot \bar{x}}{x^2 - (\bar{x})^2}$$



# Выводы

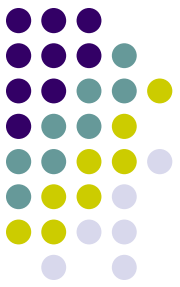
- Задача оценки степени тесноты связи между признаками решается методами корреляционного анализа.
- Если линейный коэффициент корреляции мало отличается от теоретического корреляционного отношения, то зависимость между переменными близка к линейной. Это позволяет использовать теоретическое корреляционное отношение в качестве меры линейности связи между признаками.



- Задача восстановления средних значений результативного признака по заданным значениям факторного признака решается методами регрессионного анализа.
- Использование методов корреляции и регрессии предполагает вычисление основных параметров распределения (средних величин, дисперсии).



# Вопросы для самопроверки



- Что представляют собой корреляционная связь?
- Что следует понимать под корреляцией и регрессией?
- Какие задачи решает корреляционный метод анализа?
- Что такое ложная корреляция. Каковы причины ее возникновения?
- Какими показателями измеряется теснота связи?

# Библиография



- Елисеева И.И., Юзбашев М.М. Общая теория статистики: Учебник/ Под ред. И.И. Елисеевой. – 5-е изд., перераб. и доп. – М.: Финансы и статистика, 2004.
- Кургузов В.В. Корпоративная статистика: экономико-статистическое моделирование материально-технического снабжения и сбыта. – 2006.
- Статистика для менеджеров с использованием Microsoft Excel / Д. М. Левин, Д. Стефан, Т. С. Кребиль, М. Л. Беренсон. - 4-е изд.: Пер. с англ. – М.: Издательский дом «Вильямс», 2004.
- Статистика: Учебник/ Под ред. В. С. Мхитаряна. – М.: Экономист, 2005.
- Салин В. Н. Чурилова Э. Ю. Курс теории статистики для подготовки специалистов финансово-экономического профиля: Учебник/ В. Н. Салин, Э. Ю. Чурилова – 2006.
- Практикум по теории статистики: Учеб. пособие/ Под ред. Проф. Р. А. Шмойловой. – М.: Финансы и статистика, 2004.