Медицинская статистика

Занятие 6

Корреляционный анализ

вер. 1.0 - 2004

Понятие корреляции

Регрессионный анализ отвечал на вопрос: Каков характер связи между признаками (прямолинейный, криволинейный, какой функцией эту связь можно описать).

Корреляционный анализ отвечает на вопрос: **Какова сила связи между признаками.**

Понятие корреляции отражает, главным образом, степень выраженности связи между переменными.

Понятие корреляции

Одним из подходов к корреляции является вычисление **доли объясняемой дисперсии**, т.е. доли вариабельности одного признака, зависящей от вариабельности другого.

Эта мера вычисляется по формуле: $r^2 \times 100(\%)$ (где r - коэфф. корреляции.)

Например, для коэффициента корреляции r=0,5, доля объясняемой дисперсии равна $0{,}5^2 \times 100(\%) = 25\%$

Понятие коэффициента корреляции

Основной носитель информации о корреляции - **коэффициент корреляции *(r)***

Коэффициент корреляции показывает, в какой степени изменение значения одного признака сопровождается изменением значения другого признака.

Значения коэффициента корреляции изменяются в интервалах от 1 до -1.

Крайние значения (±1) указывают на наличие линейной функциональной связи между признаками.

Ноль - на отсутствие статистической связи.

Оценка связи по силе и направлению

По направлению, связь может быть прямой и обратной, а по силе – сильной, средней и слабой. Узнать эти свойства связи позволяет коэффициент корреляции:

Сила связи	Характер связи Прямая (+) Обратная (-)
Полная	1
Сильная	1-0,7
Средняя	0,7-0,3
Слабая	0,3-0
Нет связи	0

Коэффициент корреляции вычисляют двумя способами:

1) **Параметрический метод Пирсона** (20% биомед.данных)

!Критерий согласия!

2) Непараметрические методы: **ранговой корреляции Спирмена, метод Кендалла, гамма и проч**.

$$r = \frac{\sum_{i-1}^{n} (Xi - \overline{X}) \times (Yi - \overline{Y2})}{\sqrt{\sum_{i-1}^{n} (Xi - \overline{X})^2 \times \sum_{i-1}^{n} (Yi - \overline{Y})^2}}$$

Формула коэффициента корреляции Пирсона (параметрический)

$$\rho_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

Формула коэффициента корреляции Спирмена (Ранговый, непараметрический)

Алгоритм работы с коэффициентом корреляции

1. Определение распределения данных (критерий согласия). Выбор адекватного методы вычисления к.к.

2. Вычисление коэффициента корреляции

3. Проверка статистической гипотезы о значимости коэффициента корреляции (по критерию Стьюдента)

4. Вывод о силе, направлении и достоверности связи между признаками.

1. Определение распределения данных (критерий согласия)

Если данные распределены нормально, то к ним применим параметрический метод Пирсона.

Если признаки или хотя бы один из них распределен не нормально, допустимо применение только непараметрических ранговых методов (Спирмена, Кендалла, Гамма и др.).

Проверка гипотезы о виде распределения (критерия согласия) Колмогорова-Смирнова, Лилефорса, Шапиро-Вилка.

H_0 критерия согласия: Признак распределен нормально
H_1 критерия согласия: Признак распределен не нормально

Проверка по каждому признаку!

2. Вычисление коэффициента корреляции

А) **Метод Пирсона**. См. учебник

Б) **Метод Спирмена**:

Правила присваивания рангов

Ранг наблюдения – номер, который получит наблюдение в совокупности после ранжирования (по определенному правилу).

Если отдельные наблюдения встречаются в ряду несколько раз, то каждому из них присваивается одинаковый ранг, равный среднему рангу.

Коэффициент ранговой корреляции Спирмена вычисляется (d-разность рангов):

$$\rho_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Пример вычисления коэффициента корреляции Спирмена

Температура						Пульс					Разность рангов (d)	d^2
№	Набл.	Ранжирование			Ранг	Набл.	Ранжирование			Ранг		
1	**36**	35,5	1	1	**2,5**	**90**	85	1	1	**2**	-0,5	0,25
2	**35,5**	36	2	2,5	**1**	**95**	90	2	2	**3,5**	2,5	6,25
3	**39**	36	3	2,5	**9**	**120**	95	3	3,5	**9**	0	0
4	**38**	36,6	4	4	**7**	**110**	95	4	3,5	**7**	0	0
5	**36**	37,5	5	5	**2,5**	**85**	100	5	5	**1**	-1,5	2,25
6	**40**	38	6	7	**10**	**140**	110	6	7	**10**	0	0
7	**37,5**	38	7	7	**5**	**100**	110	7	7	**5**	0	0
8	**38**	38	8	7	**7**	**110**	110	8	7	**7**	0	0
9	**36,6**	39	9	9	**4**	**95**	120	9	9	**3,5**	-0,5	0,25
10	**38**	40	10	10	**7**	**110**	140	10	10	**7**	0	0
											Сумма	9

$$\rho = 1 - \frac{6 \times 9}{10(100-1)} = 0.95$$

3. Проверка статистической гипотезы о значимости коэффициента корреляции (по критерию Стьюдента)

1. Формулировка гипотез:

H_0: Коэффициент корреляции не достоверен ($r = 0$)

H_1: Коэффициент корреляции достоверен ($r \neq 0$)

2. Уровень значимости (α=0,05)

3. Работа с критерием Стьюдента

3.1 Вручную $t_{набл} = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2}$

$t_{крит}$=табл. Критические точки t-критерия Стьюдента (Альфа и n-2 степеней свободы)

3.2 В программа СА. P-level

4. Вывод: $p > \alpha$ Нет оснований отвергать H_0

$p < \alpha$ отвергаем $H_{0,,}$ принимаем H_1

4. Вывод о силе, направлении и достоверности связи между признаками.

1. Оценка направления связи - по знаку к.к.
2. Оценка силы связи - по значению к.к.
3. Оценка достоверности связи - сравнение p и α

Сила связи	Характер связи Прямая (+) Обратная (-)
Полная	1
Сильная	1-0,7
Средняя	0,7-0,3
Слабая	0,3-0
Нет связи	0

Пример: r = 0.39 p = 0.4; r = 0.97 p = 0.04; r = -0.23 p = 0.001

Практическое задание

Изучить и оценить корреляционную связь между указанными признаками….