ПЕНЗЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Медицинский институт

Кафедра гигиены, общественного здоровья и здравоохранения

( зав. кафедрой к.м.н. А.П. Дмитриев)

**МЕТОДЫ ИЗУЧЕНИЯ КОРРЕЛЯЦИОННЫХ СВЯЗЕЙ**

Учебно-методическое пособие для студентов

(VШ семестр)

г. Пенза, 2005.

Информационный лист:

Учебно-методическое пособие “Методы изучения корреляционных связей.” подготовлено кафедрой гигиены, общественного здоровья и здравоохранения Пензенского государственного университета (заведующий кафедрой, к.м.н. Дмитриев А.П.).

В составлении принимали участие: к.м.н. Зубриянова Н.С. , Дмитриев А.П. (ответственный за подготовку Зубриянова Н.С.).

Учебно-методическое пособие подготовлено в соответствии с «Программой по общественному здоровью и здравоохранению ” для студентов лечебных факультетов высших медицинских учебных заведений”, разработанной Всероссийским учебно-научно-методическим Центром по непрерывному медицинскому и фармацевтическому образованию Минздрава России и УМЦпкп и утвержденной Руководителем департамента образовательных медицинских учреждений и кадровой политики Н.Н. Володиным в 2000 г.

Данное Учебно-методическое пособие подготовлено для студентов для самостоятельной подготовки к практическим занятиям по указанной теме.

**Тема: Методы изучения корреляционных связей**

Вопросы:

* Корреляционный анализ (основные понятия)
* Коэффициент парной корреляции
* Ранговый коэффициент (Спирмена)
* Определение тесноты связи между качественными признаками (коэффициенты Шарлье, Юла, Пирсона)

Продолжительность занятия: 4 часа

Самостоятельная работа: лабораторная работа №7

**Теоретическая часть**

# **Корреляционный анализ**

В различных областях медицины, биологии, организации здравоохранения, социально-гигиенических и клинических исследованиях проводятся статистический анализ связей, изучение закономерностей и влияющих факторов.

Существуют два вида проявления количественных взаимосвязанностей между признаками (явлениями, факторами) – функциональные и корреляционные.

При функциональных зависимостях каждому значению одной переменной величины соответствует одно вполне определенное значение другой переменной. Такие зависимости наблюдаются в математике и физике. Различные измерительные приборы основаны на функциональной зависимости (высота ртутного столбика дает однозначный ответ о температуре).

Корреляционные или статистические связи, при которых численному значению одной переменной соответствует много значений другой переменной. Пример, между ростом и весом детей существует бесспорная зависимость, но это не значит, что определенному росту строго соответствует определенный вес. В силу участия в формировании веса многих других факторов, каждому значению роста соответствует несколько значений веса, которые могут быть выражены в виде распределения.

Функциональная связь имеет место по отношению к каждому конкретному наблюдению. Корреляционная проявляется в среднем для всей совокупности наблюдений. выявления взаимодействия факторов, определение силы и направленности Практическое использование корреляционного анализа: выявление взаимодействия факторов, определение силы и направления влияния одних факторов на другие.

Следует подчеркнуть, что определение наличия связи между явлениями и факторами – дело специалистов. Статистика лишь измеряет эту связь.

Корреляционная зависимость отличается по форме связи, ее направлению и силе. Ориентировочное представление о характере зависимости между двумя изученными факторами дает графический анализ (так называемая «скэттер-диаграмма»), который позволяет рассмотреть концентрацию и рассеивание точек на пересечении координат изучаемых признаков в определенном направлении вокруг линии регрессии.

Форма связи может быть прямолинейной и криволинейной. Прямолинейная связь – равномерные изменения одного признака соответствуют равномерным изменениям второго признака при незначительных отклонениях. Криволинейная связь – равномерные изменения одного признака соответствуют неравномерным изменениям второго признака.

Направление связи может быть прямое (положительное) или обратное (отрицательное). Если с увеличением одного признака второй также увеличивается или с уменьшением одного другой тоже уменьшается, зависимость прямая, положительная. Если с увеличением одного признака другой уменьшается или с уменьшением первого признака второй увеличивается, зависимость обратная, отрицательная.

По силе связи зависимость может быть сильная (сильно выражена), средняя (умеренно выражена), слабая (слабо выражена).

## Оценка силы связи по величине коэффициента корреляции

|  |  |
| --- | --- |
| Размер связи | Характер связи |
| Прямая (+) | Обратная (-) |
| Отсутствует | 0 | 0 |
| Слабая | От 0 до +0,29 | От 0 до -0,29 |
| Средняя  | От +0,3 до +0,69 | От -0,3 до -0,69 |
| Сильная | От +0,7 до +0,99 | От -0,7 до -0,99 |
| Полная (функциональная) | +1,0 | -1,0 |

Количественная характеристика взаимосвязи изучаемых признаков может быть дана на основании вычисления показателей силы связи между ними (коэффициенты корреляции) и определения зависимости одного признака от изменения другого (коэффициент регрессии).

### **Коэффициент парной корреляции**

Коэффициент парной корреляции вычисляется по формуле:

 или 

Алгоритм расчета коэффициента парной корреляции:

1. записывают исходные данные в два вариационных ряда – x и y;
2. вычисляют среднюю арифметическую ряда x и y;
3. определяют разность между членом ряда и средними величинами;
4. перемножают разности ряда x и y между собой;
5. находят сумму перемножаемых разностей (с учетом арифметического знака);
6. возводят в квадрат каждую разность (отклонение) ряда х и у;
7. определяют сумму квадратов отклонений (разностей) для ряда х и у отдельно;
8. подставляют полученные данные в исходную формулу и вычисляют коэффициент парной корреляции.

Пример. Определить корреляционную связь между строками введения противодифтерийной сыворотки и летальностью от этого заболевания.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| День введения сыворотки (х) | Летальность (у) | dx | dy | dx2 | dy2 | dx\*dx |
| 1-й | 2,0 | -2 | -5 | 4 | 25 | 10 |
| 2-й | 3,0 | -1 | -4 | 1 | 16 | 4 |
| 3-й | 7,0 | 0 | 0 | 0 | 0 | 0 |
| 4-й | 9,0 | +1 | +2 | 1 | 4 | 2 |
| 5-й | 14,0 | +2 | +7 | 4 | 49 | 14 |
| xx = 3 | xy = 7.0 | Σdx=0 | Σdy=0 | Σdx2=10 | Σdy2=94 | Σdx\*dy =30 |



Коэффициент корреляции равен +0,98. Связь положительная, сильная. Следовательно, между сроками введения сыворотки и летальностью от дифтерии имеется очень тесная зависимость. Число больных в этом примере равно 900.

Можно определить достоверность коэффициента корреляции, вычислив его среднюю ошибку для большого числа наблюдений (n>50) по формуле:

, или при меньшем числе наблюдений: 

С достаточно большой надежностью можно утверждать, что зависимость неслучайна, если численное значение rxy превышает свою среднюю ошибку не менее чем в 3 раза.



Т.е. связь между признаками считается статистически значимой, если коэффициент корреляции превышает свою ошибку в 3 и более раз

В том случае, когда отношение коэффициента корреляции к его средней ошибки меньше 3, существование связи между изучаемыми явлениями нельзя признать доказанным.

Для малого числа наблюдений (n≤30) степень надежности коэффициента корреляции может определяться по специальной таблице. При этом число наблюдений таблицы К (число степеней свободе n) равно числу наблюдений в исследовании без двух, т.е. К = n-2. Как правило, коэффициент корреляции рассчитывается при числе коррелируемых пар не менее 5.

В медицинских и биологических исследованиях связь между признаками считается статистически значимой, если величина коэффициента корреляции больше или равна табличной при Р=0,05

Показатели оценки коэффициента корреляции при малом числе наблюдений

|  |  |
| --- | --- |
| K | P |
| 0,1 | 0,05 | 0,02 | 0,01 |
| 1 | 0,988 | 0,997 | 0,9995 | 0,99988 |
| 2 | 900 | 950 | 980 | 990 |
| 3 | 800 | 878 | 934 | 959 |
| 4 | 729 | 811 | 882 | 917 |
| 5 | 669 | 754 | 883 | 874 |
| 6 | 662 | 707 | 789 | 834 |
| 7 | 582 | 666 | 750 | 798 |
| 8 | 549 | 632 | 716 | 765 |
| 9 | 521 | 602 | 685 | 735 |
| 10 | 497 | 576 | 658 | 708 |
| 11 | 476 | 532 | 634 | 684 |
| 12 | 458 | 532 | 612 | 661 |
| 13 | 441 | 514 | 592 | 641 |
| 14 | 426 | 497 | 574 | 623 |
| 15 | 412 | 482 | 558 | 606 |
| 16 | 400 | 468 | 542 | 590 |
| 17 | 389 | 456 | 528 | 575 |
| 18 | 378 | 444 | 516 | 561 |
| 19 | 369 | 433 | 503 | 549 |
| 20 | 360 | 423 | 492 | 537 |
| 25 | 323 | 381 | 445 | 487 |
| 30 | 296 | 349 | 409 | 449 |
| 35 | 275 | 325 | 381 | 418 |
| 40 | 257 | 304 | 358 | 393 |
| 45 | 243 | 288 | 338 | 354 |
| 50 | 231 | 273 | 322 | 354 |
| 60 | 211 | 250 | 295 | 325 |
| 70 | 195 | 232 | 274 | 302 |
| 80 | 183 | 217 | 256 | 283 |
| 90 | 173 | 205 | 242 | 267 |
| 100 | 164 | 195 | 230 | 254 |

Пример. В районах изучалась зависимость между охватом населения прививками и уровнем заболеваемости. Полученный коэффициент корреляции по этим двум признакам был равен 0,81. Число наблюдений – 8 районов (пар), следовательно, К равно 6 (8-2). По таблице находим строку 6 и сравниваем полученный коэффициент. При данном числе степеней свободы (К) коэффициент корреляции превышает табличный для вероятности Р=0,05 (графа 3). Отсюда с вероятностью, большей, чем 95%, можно утверждать, что зависимость между охватом населения прививками и заболеваемостью не случайна, и эта связь сильная, т.е. чем больше процент привитых, тем меньше уровень заболеваемости.

**Ранговый коэффициент (Спирмена)**

Ранговый коэффициент корреляции более пригоден по сравнению с обычным коэффициентом для характеристики корреляций в случаях нелинейной связи и для данных, распределение которых отличается от нормального. Кроме того, данные для расчета рангового коэффициента могут быть представлены в полуколичественных измерениях. Достаточно ориентировочных данных об уровне признака. Вычисление коэффициента проводится по формуле:

.

Алгоритм расчета коэффициента корреляции рангов:

1. заменяют количественные (или полуколичественные, или качественные) признаки ряда х и у на ранги, ранжируя при этом строго от меньшей величины к большей (или строго от большей к меньшей по усмотрению исследователя);
2. определяют условные отклонения (а), т.е. разность рангов по каждой строке;
3. возводят условные отклонения в квадрат;
4. определяют сумму квадратов условных отклонений;
5. подставляют полученные данные в известную формулу и вычисляют коэффициент корреляции.

Пример. Измерить корреляцию между смертностью от рака молочной железы и рака матки (материалы смертности 5 внеевропейских стран за 1950 – 1952 гг.),

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Страна | Смертность от рака молочной железы на 100000 женщин х | Смертность от рака матки на 100000 женщин у | Порядковые номера (ранги) по размерам смертности от рака | Разность рангова | Квадрат разности рангова2 |
| молочной железы | матки  |
| Новая Зеландия | 28,6 | 14,9 | 1 | 4 | 3 | 9 |
| Австралия | 23,5 | 13,4 | 2 | 5 | 3 | 9 |
| ЮАР | 21,1 | 16,3 | 3 | 2 | 1 | 1 |
| Чили | 5,8 | 15,3 | 4 | 3 | 1 | 1 |
| Япония | 3,3 | 19,1 | 5 | 1 | 4 | 16 |
|  |  |  |  |  |  | Σа2=36 |



Поскольку вычисление коэффициента корреляции рангов обычно производится на малом числе наблюдений (число пар в рядах х, у), особое значение приобретает оценка статистической значимости (достоверности) этого коэффициента. Это тем более важно, так как ранговый коэффициент всегда менее точен, нежели парный коэффициент корреляции.

Оценка достоверности коэффициента корреляции рангов проводится разными методами в зависимости от числа наблюдений. При числе парных наблюдений, равномерном или менее 9 (n≤9), оценка значимости производится по специальной таблице. При числе наблюдений от 10 и более оценка значимости может осуществляться с помощью критерия t по формуле:

.

Используя таблицу значений t Стьюдента, при числе степеней свободы без двух (n`=n-2), сравнивают вычисленное значение с табличным. Коэффициент признается значимым при условии, если расчетное t>t0,05 табличного.

Критическое значение коэффициентов корреляции Спирмена – ρ

|  |  |  |  |
| --- | --- | --- | --- |
| n | Уровни значимости | n | Уровни значимости |
| 5% | 1% | 5% | 1% |
| 4 | 1,000 |  | 16 | 0,425 | 0,601 |
| 5 | 0,900 | 1,000 | 18 | 0,399 | 0,564 |
| 6 | 0,829 | 0,843 | 20 | 0,377 | 0,534 |
| 7 | 0,714 | 0,893 | 22 | 0,359 | 0,508 |
| 8 | 0,643 | 0,833 | 24 | 0,343 | 0,485 |
| 9 | 0,600 | 0,783 | 26 | 0,329 | 0,465 |
| 10 | 0,564 | 0,746 | 28 | 0,317 | 0,448 |
| 12 | 0,506 | 0,712 | 30 | 0,306 | 0,435 |
| 14 | 0,456 | 0,645 |  |  |  |

Коэффициент корреляции признается значимым (достоверным), если вычисленное ρ≤ρ0,05.

В рассматриваемом примере число наблюдений меньше 9 (n=5), вследствие чего оценка коэффициента проведена по таблице. При числе наблюдений, равном 5, вычисленное значение ρ=-0,80 меньше критического и поэтому нельзя с достаточной достоверностью утверждать, что между смертностью от рака молочной железы и рака матки существует какая-то зависимость.

**Определение тесноты связи между качественными признаками**

При изучении зависимости качественных признаков используется коэффициент сопряженности. Для определения тесноты связи в случае альтернативной зависимости двух сопоставляемых признаков (данные, как правило, представлены в четырехпольной таблице) коэффициенты сопряженности рассчитываются по формулам:

коэффициент контингенции Шарлье:



коэффициент ассоциации Юла (Q):

.

Алгоритм расчета:

1. шифруют через a, b, c, d четыре поля, в которых расположены исходные данные: ;
2. вычисляют последовательно произведения a\*d и b\*c;
3. рассчитывают числитель формулы;
4. определяют знаменатель формулы;
5. вычисляют коэффициент Шарлье или Юла.

Пример. Определить зависимость между методами лечения (хирургический и рентгенотерапия, только хирургический) и результатами (выздоровление или нет выздоровления).

|  |  |  |  |
| --- | --- | --- | --- |
| Метод | Выздоровление | Нет выздоровления | Всего больных |
| I | 14(a) | 8(b) | 22(a+b) |
| II | 7(c) | 9(d) | 16(c+d) |
| Всего: | 21(a+c) | 17(b+d) | 38(a+b+c+d)=N |



.

Примечание. Коэффициент ассоциации дает быструю, но ориентировочную оценку связи. Довольно точную величину коэффициент Q определяет для значений r между 1,5 и 0,5.

В тех случаях, когда качественные факторы имеют не альтернативное варьирование (четыре поля), а большее число группировок, коэффициент сопряженности вычисляется по формуле:

,

где ϕ2 (фи-квадрат) – коэффициент контингенции Пирсона.

Алгоритм расчета:

1. составляют корреляционную таблицу;
2. частоты (числа наблюдений) вписывают наверху каждой клетки таблицы;
3. частоты возводят в квадрат и результат записывают под ними;
4. квадрат частот делят на сумму числа наблюдений (частот) каждой графы таблицы (см. пример расчета: 625:46=12,7; 225:75=3 и т.д.);
5. частные от деления проставляют внизу каждой клетки таблицы;
6. определяют сумму частных от деления по каждой строке и записывают в нижней части итоговой клетки таблицы (12,7+3,0+2,0=17,7 и т.д.);
7. указанные суммы делят на соответствующие итоги строк (сумма от числа наблюдений по строке): 17,7:50=0,35; 24,5:60=0,4 и т.д.;
8. вычисляют значение ϕ2, которое является суммой полученных выше частных от деления без единицы (ϕ2=1,28-1=0,28);
9. полученное значение ϕ2 подставляют в формулу и вычисляют коэффициент сопряженности С.

Пример. Вычислить коэффициент сопряженности при измерении тесноты связи между жилищными условиями и заболеваемостью обследуемых.

|  |  |  |
| --- | --- | --- |
| Жилищные условия | Распределение обследуемых по частоте заболеваний | Итого |
| Частота болеющих | Эпизодически болеющие | Не болеющие |
| Неудовлетворительные | 25 | 15 | 10 | 50 |  |
| 252=625 | 225 | 100 |  |  |
| 12,7 | 3 | 2 | 17,7 | 0,35 |
| Удовлетворительные | 18 | 36 | 6 | 60 |  |
| 324 | 1296 | 36 |  |  |
| 6,6 | 17,3 | 0,6 | 24,5 | 0,4 |
| Хорошие | 6 | 24 | 40 | 70 |  |
| 36 | 576 | 1600 |  |  |
| 0,8 | 7,7 | 28,6 | 37,1 | 0,53 |
|  | 25+18+6=49 | 75 | 56 | 180 | 0,35+0,4+0,53=1,28 |

, отсюда .

Величина С позволяет судить о наличии средней корреляционной связи между рассматриваемыми факторами.

**Контрольные вопросы**

1. Дайте определение корреляционной зависимости, форма связи, направление связи.
2. Оценка силы связи по величине коэффициента корреляции.
3. Коэффициент парной корреляции.
4. Ранговый коэффициент Спирмена.
5. Определение тесноты связи между качественными признаками.

**ТЕСТЫ к практическому занятию по теме**

**« Методы изучения корреляционных связей»**

1. Наиболее простым методом определения степени связи между признаками является:

1) метод Спирмена

2) метод контингенции Пирсона

3) метод стандартизации 1.

2. Наиболее точным методом определения степени связи между качественными признаками является:

1) метод парной корреляции

2) метод ранговой корреляции

3) ϕ2 (фи-квадрат)

3. Корреляционная связь может быть прямой и

1) обратной

2) косвенной

4. Коэффициент корреляции, равный нулю, свидетельствует:

1) об отсутствии связи между явлениями

2) о слабой связи между явлениями

3) о слабой отрицательной связи между явлениями

5. Коэффициент корреляции, равный единице, свидетельствует:

1) о наличии функциональной связи между явлениями

2) о наличии сильной корреляционной связи между явлениями

6. Коэффициент ранговой корреляции рассчитывается при числе коррелируемых пар:

1) не менее 5

2) не менее 30

3) не менее 10

7. Связь между признаками считается статистически значимой, если величина коэффициента корреляции больше или равна табличной при:

1) Р=0,05

2) Р=0,5

3) Р=0,2

8. Связь между признаками считается статистически значимой, если коэффициент корреляции превышает свою ошибку:

1) в 3 и более раз

2) в 2 и более раза

3) в 1,5 и более раза

9. Коэффициент корреляции, равный «-0,3», свидетельствует:

1) о слабой отрицательной связи между явлениями

2) о средней отрицательной связи между явлениями

3) об отсутствии связи между явлениями

10. Корреляционная связь характеризуется соответствием:

1) нескольких значений одного признака одному значению второго признака;

2) одного значения первого признака строго определенному значению второго признака.

11. Практическое использование корреляционного анализа:

1) расчет обобщающих коэффициентов, характеризующих различные стороны каждого из изучаемых признаков;

2) сравнение степени однородности исследуемых совокупностей;

3) определение пределов возможных колебаний совокупностей;

4) выявление взаимодействия факторов, определение силы и направления влияния одних факторов на другие.

12. Корреляционный анализ используется для :

1) расчета обобщающих коэффициентов, характеризующих различные стороны каждого из изучаемых признаков;

2) сравнения степени однородности исследуемых совокупностей;

3) определение пределов возможных колебаний выборочных показателей при данном числе наблюдений;

4) выявления взаимодействия факторов, определение силы и направленности.

13. Корреляционной называется связь:

1) дающая полную характеристику совокупности по ее гомогенности, особенности распределения двух сравниваемых признаков;

2) при которой значению каждой величины одного признака соответствует несколько значений другого взаимосвязанного с ним признака;

3) при которой любому значению одного из признаков соответствует строго определенное значение другого взаимосвязанного с ним признака.

14. Функциональной называется связь:

1) при которой значению каждой величины одного признака соответствует несколько значений другого взаимосвязанного с ним признака;

2) дающая полную характеристику совокупности по ее гомогенности, особенности распределения двух сравниваемых признаков;

3) при которой любому значению одного из признаков соответствует строго определенное значение другого взаимосвязанного с ним признака.

15. Корреляционная связь определяется, как связь:

1) при которой любому значению одного из признаков соответствует строго определенное значение другого взаимосвязанного с ним признака;

2) при которой значению каждой величины одного признака соответствует несколько значений другого взаимосвязанного с ним признака;

3) дающая полную характеристику совокупности по ее гомогенности, особенности распределения двух сравниваемых признаков.

16. Расчет коэффициента ранговой корреляции используется для:

1) определения взаимосвязи между двумя меняющимися признаками;

2) установление связи между несколькими статистическими совокупностями;

3) для характеристики корреляций в случаях нелинейной связи и для данных, распределение которых отличается от нормального;

4) оценки достоверности различия двух величин.

17. Условия для расчета коэффициента ранговой корреляции:

1) для расчета используются негруппированные ряды значений двух признаков;

2) достаточно ориентировочных данных об уровне признака;

3) расчет изменения величины одного признака при изменении величины другого признака на единицу;

4) расчет производится только между количественными признаками.

18. Укажите правильную формулу для расчета коэффициента ранговой корреляции:

1)

2)

3)

4)

19. Расчет  используется для:

1) определения достоверности различия нескольких совокупностей по распределению в них какого-либо признака;

2) оценки достоверности различия двух средних величин;

3) определения взаимосвязи между двумя количественными признаками, один из которых представлен в виде интервалов значений;

4) определения взаимосвязи между двумя меняющимися количественными признаками.

20. Укажите правильную формулу для расчета коэффициента линейной корреляции (Пирсона):

1)

2)

3)

21. Значения коэффициента корреляции, превышающее табличное при Р = 95%, подтверждает статистическую:

1) достоверность;

2) недостоверность результатов.

22. Значение , превышающее табличное при Р = 95% , подтверждает статистическую:

1) достоверность;

2) недостоверность результатов.

УСТАНОВИТЕ СООТВЕТСТВИЕ

|  |  |  |
| --- | --- | --- |
| 23. Корреляционная связь: | Значение |  |
| 1. прямая сильная;
2. прямая слабая;

3) обратная слабая;1. обратная средней силы.

Одной цифре соответствует только одна буква. | А.Б.В.Г.Д.Е.Ж.З. | 0,20,9-0,4-0,121,5-1,1-2,41,1 |
| 24. Корреляционная связь: | Значение |  |
| 1. прямая слабая;

2) обратная средней силы;3) прямая сильная;1. обратная сильная.

Одной цифре соответствует только одна буква. | А.Б.В.Г.Д.Е.Ж.З. | 0,95-0,2-0,5-0,90,24-1,0-1,1-2,8 |
| 25. Корреляционная связь: | Значение |  |
| 1. прямая слабая;

2) обратная средней силы;3) прямая сильная;1. обратная слабая.

Одной цифре соответствует только одна буква. | А.Б.В.Г.Д.Е.Ж.З. | -0,51,40,1-0,2-1,40,92,0-0,95 |
|  |  |  |
| 26. Корреляционная связь: | Значение |  |
| 1. обратная слабая;

2) прямая сильная;3) обратная средней силы;1. прямая слабая.

Одной цифре соответствует только одна буква. | А.Б.В.Г.Д.Е.Ж.З. | 0,390,110,91,3-0,27-0,560,9-1,42 |

|  |  |  |
| --- | --- | --- |
| 27. Корреляционная связь: | Значение |  |
| 1. прямая сильная;
2. обратная средней силы.
3. прямая слабая;

4) обратная слабая;Одной цифре соответствует только одна буква. | А.Б.В.Г.Д.Е.Ж.З. | 1,0-0,20,6-0,650,250,9-0,82,5 |