**Знать** лекцию, параграф 1-3, страницы 331-342 учебник В.Е. Гмурман «Руководство к решению задач по теории вероятности и математической статистики»

**ВЫПОЛНИТЬ** **ЗАДАНИЕ:** законспектировать лекцию, решить задачу

1. Изучалась зависимость между объемом грудной клетки мужчин Y (см) и ростом X (см). Результаты наблюдений приведены в таблице:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| X | 162 | 164 | 179 | 172 | 182 | 188 | 168 |
| Y | 88 | 94 | 98 | 100 | 102 | 108 | 112 |

Вычислить выборочный коэффициент корреляции и оценить силу и направление связи между исследуемыми величинами. Уровень доверительной ве- роятности Р=0,95.

ВЫПОЛНЕННОЕ ДОМАШНЕЕ ЗАДАНИЕ ОТСЫЛАТЬ НА МОЮ ПОЧТУ

В ВИДЕ ФОТОГРАФИИ:

**furkalo25@yandex.ua**

**СРОКИ ВЫПОЛНЕНИЕ ЗАДАНИЕ 08.10.2020 ДО 10:00**

**Лекция№ 6**

**Тема «Элементы теории корреляции и линейной регрессии»**

Еще Гиппократ в VI в. до н. э. обратил внимание на то, что существует связь между телосложением и темпераментом людей, между строением тела и предрасположенностью к тем или иным заболеваниям. Можно привести и другие подобного рода примеры (связь погоды и уровня простудных заболеваний, связь активности Солнца и обострения сердечных, психических заболеваний). Поэтому естественно стремление использовать эти закономерности в интересах человека, придать им более или менее количественное описание. Для решения задач по анализу связей между величинами был разработан специальный статистический метод, получивший название корреляционного анализа. Корреляционный анализ

— это статистический метод количественного анализа связей, существующих между величинами, характеризующими какой-либо процесс или явление.

Различают два вида связей между явлениями или процессами: функциональная и корреляционная.

**Функциональная связь** — это такая связь, при которой определенному значению одной величины соответствует строго определенное значение другой величины. Например, какому-либо значению диаметра шара соответствует одно вполне определенное значение его объема. Функциональные связи легко обнаружить и измерить как на единичных, так и на групповых объектах

**Корреляционная связь** — это такая связь, когда определенному значению одной величины соответствует несколько значений другой величины. Пример - каждому значению роста человека соответствует целый набор массы тела. Корреляционные связи можно изучать только на групповых объектах методами математической статистики.

Корреляционная связь бывает линейной и нелинейной. **Линейная связь** характеризуется тем, что равным изменениям одной величины соответствуют равные изменения средних значений другой величины. Например, наблюдается соответствие между изменениями уровней систолического и диастолического давлений крови. Для **нелинейной связи** характерно то, что равным изменениям одной величины соответствуют неодинаковые изменения средних значений другой величины.

Корреляционная связь между величинами бывает прямой и обратной. При **прямой связи** с увеличением значений одной величины возрастает среднее другой величины. Например, с повышением температуры тела увеличивается частота пульса у большинства инфекционных больных. Вычисленный при этом коэффициент корреляции имеет положительное значение. При **обратной связи** с увеличением одной величины среднее значение другой величины уменьшается. Например, чем ниже температура воздуха в осенний период, тем выше заболеваемость детей острым бронхитом. Вычисленный при этом коэффициент корреляции имеет отрицательное значение.

Корреляция между величинами может быть представлена в виде таблицы, графика и коэффициента корреляции (или корреляционного отношения).

Таблица в данном случае состоит из двух строк, в каждой из столбцов находятся парные значения исследуемых величин, относящиеся к одному объекту наблюдения. Ранжирование в этой таблице не производится!

|  |  |  |
| --- | --- | --- |
| № | Хi | Yi |
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |
| ... |  |  |
| n |  |  |

График корреляционной связи представляет собой набор точек, координатами которых являются числовые значения величин, относящихся к одному объекту.

Таблицы и графики дают лишь общее представление о наличии и направлении связи. Но измерить и оценить статистическую достоверность этой связи можно лишь с помощью коэффициента корреляции.

**Коэффициент корреляции** — это число, которым измеряется сила и направление связи между исследуемыми величинами. Когда говорят просто о «коэффициенте корреляции», почти всегда имеют ввиду коэффициент корреляции Пирсона. Расчет коэффициента корреляции Пирсона возможен только в том случае, если связь между переменными носит линейный характер и переменные величины подчиняются нормальному закону распределения. В некоторых случаях, когда связь нелинейная или используются качественные признаки, применяют коэффициент ранговой корреляции Спирмена, который относится к категории непараметрических методов и формулу для которого мы здесь не приводим.

Коэффициент корреляции Пирсона *r* для величин *X* и *Y* находится по фор-

муле:

*r*  *n*   *X i*  *Yi*   *X i*  *Yi* ,

*(n*   *X* 2  *(*  *X )*2 *)*  *(n*  *Y* 2  *(* *Y )*2 )

*i*

*i*

*i*

*i*

где n – число коррелирующих пар,

Х – значение независимой переменной, Y – значение зависимой переменной,

Положительные значения коэффициента корреляции указывают на прямую связь между исследуемыми величинами, а отрицательные - на обратную.

# Свойства коэффициента корреляции:

1. Коэффициент корреляции принимает значения на отрезке [- 1; 1], т. е.

– 1 ≤ r ≤ 1.

1. При r = ±1 корреляционная связь представляет линейную функциональную зависимость.
2. При r = 0 линейная корреляционная связь отсутствует.
3. Если все значения переменных увеличить (уменьшить) в одно и то же число раз, то величина коэффициента корреляции не изменится.

Числовые значения его, близкие к единице, говорят о наличии тесной связи между рассматриваемыми величинами; близкие к нулю - о слабой связи. В зависимости от величины *r* для обозначения силы связи приняты следующие названия:

|  |  |
| --- | --- |
| величина коэффициента корреля-ции по абсолютной величине | название связи |
| r = 0 | нет связи |
| 0 < r < |0,3| | слабая |
| |0,3| ≤ r < |0,7| | средняя |
| |0,7| ≤ r < |1| | сильная |
| r = 1 | функциональная |

Коэффициент корреляции так же, как и среднее арифметическое, изменяется от выборки к выборке при повторных исследованиях. Мерой изменчивости коэффициента корреляции служит **ошибка коэффициента корреляции** *m*, находимая по следующей формуле:

 *m*  ,

1  *r* 2

*n*  2

где *n* - число коррелирующих пар случайных величин, r – коэффициент корреляции.

Зная ошибку коэффициента корреляции, можно вычислить **критерий достоверности** *t* для него по формуле:

1  *r*2

*n*  2

*r*

*t*  *m*

 *r* 

Вертикальные черточки в формуле означают, что полученное значение берется по модулю, т.е. всегда является положительным числом.

Вычисленное значение критерия достоверности *t* сравнивается со значением критерия Стьюдента *tst* , находимом в таблице соответствующих коэффициентов на пересечении строки с номером k = n-2 и столбца с выбранным уровнем доверительной вероятности Р. Если t>*tst*, делается вывод о наличии корреляционной связи между величинами. Если же t<*tst*, то связь недостоверна, случайна.

Коэффициент корреляции позволяет измерить степень сопряженности между величинами, определить силу и направление существующей между ними связи. Но корреляционный анализ не может ответить на вопрос, насколько будет изменяться переменная величина, если связанная с ней корреляционно другая будет возрастать или убывать. Подобного рода задачи решаются методом регрессионного анализа.

**Регрессионный анализ** — это статистический метод, устанавливающий количественно форму зависимости двух случайных величин, между которыми существует корреляционная связь.

Основное понятие регрессионного анализа — это понятие регрессии.

**Регрессия** - функция, позволяющая по значению одной переменной величины определить средние значения другой величины, связанной с первой корреляционно. *Регрессия бывает линейная и нелинейная*. В случае линейной регрессии выполняется следующее условие: для любого значения одной величины при ее изменении на единицу среднее значение другой величины изменяется на одно и то же число. Для нелинейной регрессии это условие не выполняется. С помощью линейной регрессии можно выяснить, как количественно меняется одна величина при изменении другой величины на единицу. Для определения размера этого изменения применяется специальный коэффициент - **коэффициент регрессии**, равный числу, на которую в среднем изменяется переменная величина при изменении другой, связанной с ней корреляционно, на единицу.

С помощью коэффициента регрессии можно без специальных измерений

определить значение одной величины, зная значение другой. Для этих целей служит **уравнение линейной регрессии**, которое имеет следующий вид:

*Y*  *a*  *b*  *X* , где Х – значение независимой переменной,

Y – значение зависимой переменной,

*a, b* – параметры уравнения (*а* – коэффициент сдвига,

*b* – коэффициент наклона или коэффициент регрессии).

Параметры уравнения *a* и *b* находят, используя метод наименьших квадра-

тов.

Коэффициент сдвига:

*a*   *Yi* *Xi2*   *Xi*   *Xi*  *Yi* .

*n*   *X* 2  ( *X* )2

*i*

*i*

Коэффициент регрессии:

*b*  *n*   *Xi*  *Yi*   *Xi*  *Yi* .

*n*   *X* 2  ( *X* )2

*i*

*i*

Вычисления можно упростить, если сначала вычислить коэффициент регрессии *b*, а затем коэффициент сдвига вычислить по формуле:

*a*  *Y*  *b*  *X* ,

где *X* и *Y* - средние арифметические для переменных X и Y.

Если мы будем знать значения коэффициентов *a* и *b*, то подставляя в уравнение линейной регрессии конкретное значение независимой переменной Х, всегда можно рассчитать соответствующее ей значение зависимой переменной Y.

Меру колебания случайной величины вокруг линии регрессии характеризует

 - **среднее квадратическое отклонение регрессии**:

*b*  * y* 

1 *r*2

где * y*

- среднее квадратическое отклонение параметра Y.

* y*  

Зная коэффициент регрессии и используя уравнение регрессии и среднее квадратическое отклонение регрессии, можно на графике построить линию регрессии. Для этого необходимо выбрать два значения случайной величины *Х* (желательно *Xmin* и *Xmax*) и рассчитать по уравнению регрессии соответствующие им значения *Y*. Затем нанести координаты этих двух точек на график и провести че- рез них прямую, которая и будет линией регрессии.

*Y* 2  (*Y* )2



*Y*

2

*i*





*Y*



2

*n*

 



*i* 



*n*





На этом же графике можно указать **доверительную зону регрессии**, т.е. зону, в которую попадает вполне определенная, заранее заданная доля случайных величин данной совокупности. Зона будет образована двумя параллельными линиями, идущими одна выше, а другая ниже линии регрессии и отстоящими от нее на определенном расстоянии, называемом доверительным интервалом. Для расчета величины доверительного интервала по заданной доле используются специальные формулы, которые здесь не приводятся. Результаты расчета доверительного интервала для некоторых случаев приведены в таблице:

|  |  |
| --- | --- |
| Доля величин, попадающих в дове- рительную зону регрессии | Ширина довери- тельного интер- вала |
| 100% | 3 |
| 95% | 2 |
| 68% |  |
| 50% | 2/3 |

Такой график позволяет, например, задавшись долей величин, входящих в норму, оценить, попадает или нет конкретный, данный объект наблюдения в норму или нет, а также решать и другие задачи.

Ниже приведен пример подобного графика, на котором показана линия регрессии, отображающая связь между ростом человека и длиной мизинца. Пунктирные линии выше и ниже линии регрессии образуют доверительную зону регрессии. Ширина доверительного интервала взята равной *2*. Это означает, что в доверительную зону регрессии попадает 95% величин исследуемой совокупности.

120

Длина мизинца, мм





Рост человека, см

100

80

60

150 160 170 180 190 200

# ПРИМЕРЫ РЕШЕНИЯ ЗАДАЧ.

**Задача 1.** Изучалась зависимость между систолическим давлением Y мужчин в начальной стадии шока и возрастом X*.* Результаты наблюдений приведены в таблице:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *X* | 68 | 37 | 50 | 53 | 75 | 66 | 52 | 65 | 74 | 65 | 54 |
| *Y* | 114 | 149 | 146 | 141 | 114 | 112 | 124 | 105 | 141 | 120 | 124 |

Вычислить выборочный коэффициент корреляции и оценить силу и направление связи между исследуемыми величинами. Составить уравнение линейной регрессии.

Уровень доверительной вероятности Р=0,95.

# Решение:

Составим расчетную таблицу:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № | Хi | Yi | Хi·Yi | Хi 2 | Yi 2 |
| 1 | 68 | 114 | 7752 | 4624 | 12996 |
| 2 | 37 | 149 | 5513 | 1369 | 22201 |
| 3 | 50 | 146 | 7300 | 2500 | 21316 |
| 4 | 53 | 141 | 7473 | 2809 | 19881 |
| 5 | 75 | 114 | 8550 | 5625 | 12996 |
| 6 | 66 | 112 | 7392 | 4356 | 12544 |
| 7 | 52 | 124 | 6448 | 2704 | 15376 |
| 8 | 65 | 105 | 6825 | 4225 | 11025 |
| 9 | 74 | 141 | 10434 | 5476 | 19881 |
| 10 | 65 | 120 | 7800 | 4225 | 14400 |
| 11 | 54 | 124 | 6696 | 2916 | 15376 |
| ∑ | 659 | 1390 | 82183 | 40829 | 177992 |

лицы.

1. Вычислить произведение Хi·Yi.
2. Вычислить Хi 2.
3. Вычислить Yi 2.
4. Найти суммы значений, т. е. заполнить последнюю строку расчетной таб-
5. Вычислить коэффициент корреляции:

*r*  *n*   *Xi*  *Yi*   *Xi*  *Yi* 

*(n*   *X* 2  *(*  *X )*2 *)*  *(n*  *Y* 2  *(* *Y* )2 )

*i*

*i*

*i*

*i*

 11 82183 659 1390

(11 40829 6592 )  (11177992 13902 )

  11997

19570,346

 0,613

1. По формуле ента корреляции:

*t* *r* 

вычислить критерий достоверности коэффици-

*t*  0,613 11  2  2,33.

*n*  2

1  *r* 2

1  0,6132

Критерий Стьюдента для k = n-2 = 11-2 = 9 и Р=0.95 *tst* = 2,26. 7.Делаем вывод:

Так как *t > tst*, то с вероятностью Р=0,95 можно утверждать, что между систолическим давлением мужчин в начальной стадии шока и возрастом существует средняя и обратная корреляционная связь.

8. Уравнение линейной регрессии имеет следующий вид:

*Y*  *a*  *b*  *X* , где Х – значение независимой переменной,

Y – значение зависимой переменной,

*a, b* – параметры уравнения (*а* – коэффициент сдвига,

*b* – коэффициент наклона или коэффициент регрессии). а) Вычислить средние арифметические:

*X*   *Xi*

*n*

 659  59,91

11

*Y*  *Yi*

*n*

 1390  126,36

11

б) Вычислить значение коэффициента регрессии по формуле:

*b*  *n*   *Xi*  *Yi*   *Xi*  *Yi*

*n*   *X* 2  ( *X* )2

*i*

*i*

Суммы для расчета берутся из таблицы, нижняя строка.

Коэффициент регрессии:

*b*  *n*   *Xi*  *Yi*   *Xi*  *Yi*

 11 82183 6591390  11997  0,81.

*n*   *X* 2  ( *X* )2

*i*

*i*

11 40829 6592

14838

в) Вычислить коэффициент сдвига:

*a*  *Y*  *b*  *X*  126,36  0,81 59,91  174,89 ,

Уравнение регрессии: Y = 174,98 - 0.81 . Х.

# Вывод:

При изменении возраста мужчин на 1 год систолическое давлением в начальной стадии шока изменяется в среднем на величину, равную коэффициенту регрессии *b* = -0,81 ед.

ПРИЛОЖЕНИЕ

# СТАНДАРТНЫЕ ЗНАЧЕНИЯ КРИТЕРИЯ СТЬЮДЕНТА

|  |  |  |  |
| --- | --- | --- | --- |
| К | УРОВНИ ДОВЕРИТ. ВЕРОЯТНОСТИ0,95 0,99 0,999 | К | УРОВНИ ДОВЕРИТ. ВЕРОЯТНОСТИ0,95 0,99 0,999 |
| 1 | 12,71 | 63,66 | 636,62 | 18 | 2,10 | 2,88 | 3,92 |
| 2 | 4,30 | 9,92 | 31,60 | 19 | 2,09 | 2,86 | 3,88 |
| 3 | 3,18 | 5,84 | 12,94 | 20 | 2,09 | 2,85 | 3,85 |
| 4 | 2,78 | 4,60 | 8,61 | 21 | 2,08 | 2,83 | 3,82 |
| 5 | 2,57 | 4,03 | 6,86 | 22 | 2,07 | 2,82 | 3,79 |
| 6 | 2,45 | 3,71 | 5,96 | 23 | 2,07 | 2,81 | 3,77 |
| 7 | 2,36 | 3,50 | 5,40 | 24 | 2,06 | 2,80 | 3,74 |
| 8 | 2,31 | 3,36 | 5,04 | 25 | 2,06 | 2,79 | 3,72 |
| 9 | 2,26 | 3,25 | 4,78 | 26 | 2,06 | 2,78 | 3,71 |
| 10 | 2,23 | 3,17 | 4,59 | 27 | 2,05 | 2,77 | 3,69 |
| 11 | 2,20 | 3,11 | 4,49 | 28 | 2,05 | 2,76 | 3,66 |
| 12 | 2,18 | 3,05 | 4,32 | 29 | 2,05 | 2,76 | 3,66 |
| 13 | 2,16 | 3,01 | 4,22 | 30 | 2,04 | 2,75 | 3,65 |
| 14 | 2,14 | 2,98 | 4,14 | 40 | 2,02 | 2,70 | 3,55 |
| 15 | 2,13 | 2,95 | 4,07 | 60 | 2,00 | 2,66 | 3,46 |
| 16 | 2,12 | 2,92 | 4,01 | 120 | 1,98 | 2,62 | 3,37 |
| 17 | 2,11 | 2,90 | 3,96 |  | 1,96 | 2,58 | 3,29 |

К - число степеней свободы.

К = N1 + N2 - 2 (для критерия достоверности разности средних арифметических двух выборок).

К = N - 1 (для одной выборки).