04 & 05 - Упражнение

Изследване на връзки. Корелация. Асоциация

ас.д-р Костадин Костадинов

Вместо увод.

Хората сме податливи на бързи заключения. Често мозъкът ни "заблуждава" да мислим за две събития като свързани. По този начин по-лесно си обясняваме света около нас. Това е характерно и за медицината. В статистиката тази връзки са от особено значение. Наличието на статистическа връзка е едно от условията за да наречем даден фактор "рисков" за дадено заболяване. Статистическата връзка е важен аргумент и когато се опитваме да докажем, че едно лекарство е ефективно.

В курса по статистика, ще се спрем на най-опростеният вариант за връзка - тази между само две променливи.[1]

Когато изследваме връзката между две количествени променливи (тегло, ръст, кръвно налягане и т.н) използваме термина **корелация**, а когато изследваме връзката между две качествени променливи (пол, заболеваемост, тютюнопушене, диагноза) използваме термина **асоциация**.

[1] Например - връзката между тютюнопушенето и заболеваемостта от рак на белия дроб или връзката между кръвното налягане и теглото на човека.

Природа на връзката

- Статистическа връзка не означава непременно **причинно следствена връзка!** С други думи, статистическата връзка може да бъде установена, дори и без да има естествена причинно-следствена такава такава.[2]

[2] Замъгляването може да бъде отстранено, чрез стандартизация.

- Статистическата връзка, може да бъде “замъглена”. [3]
- Статистическата връзка не бива да се екстраполира безразсъдно.

[3] Замъгляването може да бъде отстранено, чрез стандартизация.

Екстраполация

Екстраполацията е математически метод за намиране на нови стойности на търсена функция или връзка, извън множеството известни нейни стойности, служещи за построяването (установяването) и.

Ето един пример. Да приемем, че с увеличаването на теглото ръста също расте (и обратно с увеличаването на ръста, се увеличава теглото). Тази връзка е очевидна при децата например, с растежа те наддават както на височина, така и на тегло. Същата връзка обаче изчезва при възрастните. Растежът на височина е завършил, а теглото може да се увеличава (понякога и безконтролно). Тоест, ако пренесем връзката между тегло и ръст наблюдавана при децата върху възрастните, ще допуснем сериозна грешка.

Асоциация

Тестовете за асоциация се използват за установяване на връзка между качествени променливи.[4]

Нека разгледаме един лесен пример, за да обясним теста за асоциация.

Нека си представим, че екип от лекари е направил проучване, в което тества три вида лекарства за лечение на КОВИД-19. Лекарствата ще наречем “А”, “В и”С”. В проучването са включени 500 пациенти. С лечение А са лекувани 150 пациенти, с лечение В са лекувани 250 пациенти, а с лечение С са лекувани 100 пациента. След 1 месец тестване, вече имаме някакви резултати. 180 от пациентите са починали, 220 са с частично подобрение, не са хоспитализирани, но не са и оздравели напълно. От всички 500 пациента 100 са напълно оздравели.

[4] Например, същества ли връзка между тютюнопушенето и рака на белия дроб или между употребата на кафе и сърдечните аритмии.

Преди да видим точно колко пациенти спрямо терапията са със съответните резултати, нека да нанесем данните с които разполагаме в таблица.

лечение	починали	частично оздравели	напълно оздравели	$\sum$ по редове
л-ство А	клетка а?	клетка г?	клетка ж?	150
л-ство В	клетка б?	клетка д?	клетка з?	250
л-ство С	клетка в?	клетка е?	клетка и?	100
$\sum$ по колони	180	220	100	**500**

Дефиниране на хипотези

Нулевата хипотеза твърди, че не съществува асоциация (връзка) между вида на терапията и резултата от лечението. С други думи, каквото и хапче да даваме на тези пациенти, резултата може да е един от трите възможности с еднаква вероятност. Лечението е независимо от резултата.

Алтернативната хипотеза твърди, че същества асоциация между лекарството и резултата от лечението. С други думи, вида на лекарството е свързано с резултата от лечението.

Ниво на грешка

Идеята на теста за асоциация е много подобна със всеки друг статистически тест. Накрая трябва да получим число, което можем да преобразуваме в вероятност то да е случайно. В случая това число се нарича χ^2. Този коефициент е подобен на t статистиката. След като получим каква е стойността на χ^2 имаме таблица, с която можем да установим каква е стойността на p [5].

[5] p е вероятността да сме получили тази или по-висока стойност в резултат на случайност! В медицината е прието, тази стойност на p да е по малко от 0,05. Това е нашият риск от грешка α

Както при тестването на хипотези до сега, отново трябва да установим при какво ниво на p ще отхвърлим нулевата хипотеза. С други думи, колко малка трябва да е вероятността, нашите резултати да са случайни за да отхвърлим нулевата хипотеза.

Изчисляване на теоретични честоти.

Може би сте забелязали по горе, че имаме доста "?"- знаци в клетките. Нека предположим, че все още не знаем, точно колко пациенти, лекувани с конкретните A, B и C медикаменти са оздравели, починали и частично оздравели. Тези данни са все още скрити за нас. Въпреки това, ние разполагаме с **нашата нулева** хипотеза. В основа на нея, можем да установим каква е **теоретичната честота** в тези клетки.

Теоретична честота е тази стойност, която трябва да запишем в клетките, ако не съществува никаква връзка между терапията и резултата.

Как можем да установим тази стойност? Разбира се имаме формула:[6]

$$f_t cell = \frac{\sum_{columns} \cdot \sum_{rows}}{\sum_{total}}$$

С други думи, ако очакваме да няма никаква връзка между лечението и резултата, пациентите лекувани с терапия A и са починали, трябва да бъдат 54. Тази стойност установихме, като умножихме сумата по редове за лечение "A" 150 с сумата за "починали" 180, а след това разделим на броя на участниците.[7]

Нека сега напълним цялата таблица с изчислените теоретични честоти.Това са **очакваните резултати**, ако нямаше никаква връзка между лечението и резултата.

[6] Положителна връзка (позитивна корелация) се наблюдава, тогава когато увеличаването на единия признак се съпровожда с увеличаването на другия (увеличаването на храната от Макдоналдс е свързано с увеличение на коремната обиколка). Негативна корелация (отрицателна връзка) се наблюдава, когато увеличението в една променлива е свързано с намаление в друга (увеличението на храната от Макдоналдс е свързано с намаление на парите в джоба)

[7] Линейна е тази зависимост *връзка* в която промяната на признаците е в линеина зависимост. Например - увеличението на приема с 1 литър вода води до увеличени с 1 литър в отделените течности.

лечение	починали	частично оздравели	напълно оздравели	$\sum$ по редове
л-ство А	*54*	*66*	*30*	150
л-ствоо В	*90*	*110*	*50*	250
л-ство С	*36*	*44*	*20*	100
$\sum$ по колони	180	220	100	**500**

Сега е момента да разберем реалните данни, така наречени обсервирани или наблюдавани честоти f_o.

лечение	починали	частично оздравели	напълно оздравели	$\sum$ по редове
л-ство А	65	*70*	*15*	150
л-ство В	*100*	*110*	*40*	250
л-ство С	*15*	*40*	*45*	100
$\sum$ по колони	180	220	100	**500**

Сега, вече можем да установим каква е разликата между нашите очаквания (ако нямаше никаква връзка) и реалността. [8]

[8] Например в ситуацията, в която очакваме да няма асоциация между лекарството и резултата, очаквахме броя на пациентите с лекарство А, които са починали да са 54, докато в реалността те са 10 повече. За пациентите, които са използвали терапия С и са оздравели, очакванията на нулевата хипотеза са 20, докато истинските бройки са 45.

Очевидно има разлика между очакванията на нулевата хипотеза и реалните данни. Но как да оценим точно колко е това "несъответствие". За целта имаме формула - това всъщност е χ^2. Колкото е по-голяма стойността на χ^2 толкова това "несъответствие" на реалността с очакванията на нулевата хипотеза е по-голяма. Колкото по-малка е стойността на χ^2 толкова по-малко е това "несъответствие".

Нека разгледаме формулата:

$$\chi^2 = \sum \frac{(f_0 - f_t)^2}{f_t}$$

За да изчислим стойността на χ^2 е необходимо за всяка една клетка да изчислим разликата между очакваната и теоретичната честота. След това повдигаме тази разлика на

квадрат. Разделяме полученето на теоретичната честотата. Правим това за всички клетки и накрая събираме получените стойности

$$\chi^2 = \frac{(65–54)^2}{54}+\frac{(70–66)^2}{66}+\frac{(15–30)^2}{30}+\frac{(100–90)^2}{90}+\frac{(110–110)^2}{110}+$$

$$\frac{(40–50)^2)}{50} + \frac{(15–36)^2}{36} + \frac{(40–44)^2)}{44} + \frac{(45–20)^2}{20}$$

$$\chi^2 = 2,2 + 0,2 + 7,5 + 1,1 + 0 + 2 + 12,3 + 0,4 + 31,3 = 57$$

Вече имаме финалната оценка на "несъответствието" между това което нулевата хипотеза твърди за лекарството и резултата и това, което ние действително сме наблюдавали.

Определяне на p-стойност

Следва да определим, колко вероятно е, това несъответствие да е чиста случайност. За да го направим, отново имаме таблица подобна на t статистиката. В тази таблица за всяка стойност на χ^2 имаме вероятност с която да сме я получили. Отново важи правилото и за t статистиката. Колкото по-голямата е стойността на χ^2 , толкова по-малко вероятно, да я наблюдаваме. Но преди да потърсим в таблицата е необходимо да знаем още една стойност. Това е стойността на df.

df представлява степен на "свобода". Това е число, което зависи от размера на нашата таблица - колко колони и колко редове има тя.

$$df = (N_{columns} - 1) \cdot (N_{rows} - 1)$$

[9]

[9] В нашия случай имаме 3 колони и 3 реда. Това е таблица 3x3. Степените на свобода за тази таблица са

$$df = (3-1) \cdot (3-1) = 2 \cdot 2 = 4$$

Сега вече можем да погледнем таблицата за χ^2. Нас ни интересува само един ред от нея - реда в който степените на свобода са 4.[10]

df	$p = 0.1$	$p = 0.05$	$p = 0.01$
4	χ^2 =7,779	χ^2 =9,488	χ^2 =13,27

[10] Най-силна корелация имаме в случаите, когато корелационния коефициент е със стойност -1 и +1. Когато корелационен коефициент равен на 0-ла **НЕ СЕ** наблюдава корелация.

Както виждаме, с увеличаването на стойността на χ^2, вероятността p спада. Нашата стойност $\chi^2 = 57$ е по-голяма от последната която виждаме в таблицата 13,27, това означава, че вероятността p е по-малка от 0,01.

Заключение

Отхвърляме нулевата хипотеза и приемаме алтернативната. Има статистически значима **връзка (зависимост)** между избраното лечение и резултата от него при COVID пациенти. Избраното лечение е фактор за степента на възстановяване.

Малко повече.

- За да прилагаме този тест е необходимо да имаме поне 5 наблюдения във всяка една клетка. При по-малка, математическия апарат е подвеждащ и не можем да имаме доверие на получената стойност за p.
- Асоциацията, подобно на корелацията е вид статистическа връзка. Това не означава, че е непременно причинно следствена. Действително установихме, че вида на лекарството оказва влияние върху резултат, но това не е достатъчно, да посочим че именно определено лекарство е причината за по-добър или по-лош резултат.
- Когато работим с таблици 2x2 (например ваксиниран/неваксиниран и инфектиран/здрав), можем да използваме корелационен анализ. Въпреки, че в изложението дефинирахме корелацията като подходяща само за количествени признаци, в тези случай, когато имаме качествени

признаци със само две възможни стойности тя е приложима. Коефициентът, който се изчислява се означава с ϕ.

Ето един пример в който можем да използваме този коефициент:

Ваксина	Имунен отговор	Без имунен отговор
Нова	“a”	“c”
	90	10
Стара	“b”	“d”
	160	90

$$\phi = \frac{(a \cdot d) - (b \cdot c)}{(a + b) \cdot (c + d) \cdot (a + c) \cdot (b + d)}$$

Корелация

> ⚠ Корелационния анализ
>
> Статистически метод, използван за изследване на взаимовръзка между две количествени променливи

Съществуват множество методики с които се анализа една корелационна връзка. Всеки от тях има свои предимства и недостатъци. Въпреки това всички те споделят някои основни изисквания, за да може корелационният анализ да бъде завършен:

- Да се оцени **силата на връзката**.
- Да се оцени **посоката** на връзката. [11]
- Корелационният анализ следва да определи “формата” на връзката. Има много видове връзки. Но за улесним анализа, в курса по статистика ще се спрем единствено на “линейна връзка” [12].

[11] Положителна връзка (позитивна корелация) се наблюдава, тогава когато увеличаването на единия признак се съпровожда с увеличаването на другия. Негативна корелация (отрицателна връзка) се наблюдава, когато увеличението в една променлива е свързано с намаление в друга.

[12] Линейна е тази зависимост *връзка* в която промяната на признаците е в линеина зависимост. Например - увеличението на приема с 1 литър вода води до увеличени с 1 литър в отделените течности.

Корелационен коефициент.

Корелационният коефициент е число в диапазона -1 до +1. Знакът -/+ показва **посоката** на корелацията. Отрицателен корелационен коефициент означава отрицателна (негативна) връзка. Корелационният коефициент отразява **силата** на връзката. Колкото по отдалечена е стойността на корелационния коефициент от 0 -лата, толкова **по-силна** е установената връзка. Същества и скала, с която да се оцени каква е силата на връзката според този коефициент.

ρ	сила на връзката
1	максимално силна връзка
0,67 до 1	силна връзка
0, 34 до 0, 66	умерена връзка
0 до 0,33	слаба връзка
0	липсва връзка
0 до - 0,33	слаба връзка
- 0,34 до -0,66	умерена връзка
- 0,67 до - 1	силна връзка
-1	максимално силна връзка

Оценка на силата на връзката

Корелационния коефициент на Пирсън, се означава с гръцката буква ρ или с латинската r. Този коефициент е много зависим от разпределението на променливите, за които търсим корелация, както и от формата на тази връзката. Коефициентът на Пирсън е подходящ, когато променливите са с нормално разпределение, а връзката е линейна.

Малко математика

Корелационният коефициент се изчислява с помощта на следната формула:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Нека предположим че имаме само 3 наблюдения (за да улесним пресмятанията)

име	Ръст	Тегло
Мария	160	50
Иван	170	90
Чавдар	180	100
средна аритметична	*170*	*80*

Тази сложна формула,може да се разгледа на части:

- $(x - \bar{x})$ представлява отклонението на всяка една точка от средната аритметчична за променливата x - ръст
- $(y - \bar{y})$ представлява отклонението на всяка една точка от средната аритметчична за променливата y - височина [13]

[13] Отклонението за ръста на Мария е -10 см, защото е с 10 см. по ниска от средната за групата. Отклонението на Чавдар е $+10$. За променливата y отклонението за Мария е -30, защото тя е 30 кг. по-лека от средната за групата, за Чавдар е $+10$ кг.

Нека нанесем тези отклонения в таблицата:

име	Ръст	Тегло	$(x - \bar{x})$	$(y - \bar{y})$
Мария	160	50	-10	-30
Иван	170	90	0	10
Чавдар	180	100	10	20
	$\bar{x} = 170$	$\bar{y} = 80$		

- След като сме направили това, можем да преминем към изчисляване на т.н сума на кръстосания продукт. С други думи, за всеки участник (i), трябва да умножим девиациите от средните аритметични и да ги сумираме.

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

.

За Мария например, това означава да умножим -10 (девиацията на ръста от средния ръст за групата) с -30 (девиацията на теглото и спрямо теглото на групата). След като направим това за всички участници трябва да сумираме тези продукти. Така, ще получим числителя на формулата. Нека да видим как изглежда това таблично

име	Ръст	Тегло	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x}) \cdot (y - \bar{y})$
Мария	160	50	-10	-30	300
Иван	170	90	0	10	0
Чавдар	180	100	10	20	200
	$\bar{x} = 170$	$\bar{y} = 80$			$\sum = 500$

След като вече имаме числителя, нека погледнем, какво е нужно за знаменател в тази сложна формула.

$$\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

Основен елемент тук е да изчислим $(x-\bar{x})^2$ и $(y-\bar{y})^2$. Ние, вече имаме тези девиации. Сега единствено трябва да ги повдигнем на квадрат.

име	Ръст	Тегло	$(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$
Мария	160	50	-10	100	-30	900
Иван	170	90	0	0	10	100
Чавдар	180	100	10	100	20	400
	$\bar{x}=170$	$\bar{y}=80$		$\sum=200$		$\sum$=1400

Вече сме готови за заместим в голямата формула, числата които ни интересуват.

$$\rho=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$\rho=\frac{500}{\sqrt{200}\cdot\sqrt{400}}$$

$$\rho=\frac{500}{14,1\cdot 37,42}=\frac{500}{529}=0,94$$

В крайна сметка, сред тези трима студенти се наблюдава силна корелация между ръста и теглото - увеличеното тегло (от Мария до Чавдар) се свързва и с увеличен ръст (отново в същият ред).

Този коефициент е валиден за тези трима студенти, но дали е валиден за генералната съвкупност?

Можем да установим това, чрез тестове на хипотеза. В случая:

Нулевата хипотеза твърди, че популационния регресионен коефициент е нула. С други думи не съществува корелация между теглото и ръста.

Алтернативната хипотеза твърди, че че в дейсвителност, има корелация и тя не е равна на 0-ла.

За да тестваме това използваме тест подобен на T теста, който вече знаем. В крайна сметка получаваме число t статистика, което ползваме за да установим, колко е вероятно да сме получили този резултат случайно (p). В случай, че тази вероятност е под 0,05 (5%) можем да отхърлим нулевата хипотеза.

Малко повече

Освен корелацията на Пирсън, съществуват и различни видове други корелационни коефициенти (Кнедау Тау, Гудман-Крускал,Somepc, Спирман, Юл, Рангово-бисериални и др.) всеки с различна формула и носещ по-различна информация от коефициента, които видяхме по-рано. Въпреки това, няма да се спираме на тях в този курс, за да ви предпазим от страничния ефекти на математическите формули да изместват химическите.