

04 & 05 – Упражнение 2023/2024

Изследване на връзки. Корелация. Асоциация

ас.г-р Костагин Костадинов

Увод

Често при представяне на статистически данни сме податливи на бързи заключения за свързаността между две или повече явления. Наличието на статистическа връзка е особено важно за медицината, тъй като тя е основна при определянето на конкретен фактор за "рисков" за развитието на заболяване. Въпреки това способността ни да изграждаме "причинно-следствени" вериги понякога ни "заблуждава". Човешкият мозък е склонен да мисли за две явления като свързани, дори и в действителност такава връзка да липсва.

Не бива да се бърка статистическа и **причинно следствена свързка**. С други думи, резултатите от статистическия анализ сами по себе си **не могат** да се разглеждат като изчерпателно доказателство на тезата за причинната обусловеност¹.

Вероятността дадена връзка да бъде причинно-следствена е по-голяма:

1. Когато тя е **по-силна** и характеризира се с високи стойности за статистическите коефициенти.
2. Когато тя е **правдоподобна** от биомедицинска гледна точка. Научните (теоретични, експериментални) данни и доказателства за механизма на взаимодействията са основни аргументи в полза на причинността. Изясняването на правдоподобността на връзката се предопределя от нивото на достигнатото познание за явленията².
3. Когато тя се потвърждава и в **други сродни проучвания**, тоест има устойчив характер.
4. Когато **времевата последователност** на причината и следствието съответства на специфичния механизъм на взаимодействията между явленията³.

¹ В повечето случаи, взаимоотношенията в областта на биологията, медицината и здравеопазването са твърде сложни, за да се обяснят единствено със статистическите коефициенти.

² Липсата на научни знания за механизма на - взаимоотношенията обаче не означава, че причинно-следственият характер на връзката трябва да бъде отхвърлен.

³ Проверката по този критерий може да не бъде толкова лесна и проста задача, колкото изглежда на пръв поглед. Понякога трябва да се преодолеят редица трудности, за да се маркират времевите моменти на причинно-следствените промени

Основни понятия

В настоящия курс са разгледани най-улеснените варианти на статистически връзки – тази между две променливи. Връзката между два **количествени** признака (нп. тегло, ръст, кръвно налягане) се означава като **корелация** и се изследва с помощта на корелационен коефициент. Връзката между два **качествени** признака (нп. пол, заболяемост, тютюнопушене, диагноза) се изследва с помощта на теста за **асоциация**.

Статистическата връзка се изследва в **обеа** на наличните данни от извадката. Предположението и в по-широк мащаб и при стойности извън наличните в извадката е пример за екстраполация ⁴.



Екстраполация

Екстраполацията е математически метод за намиране на нови стойности на конкретна променлива, следствие на установена статистическа връзка, които са извън интервала при нейното построяване.

⁴ Да приемем, че в една извадка (предимно с деца) сме наблюдавали връзка между увеличаването на теглото и увеличаването на ръста. Същата връзка обаче не бива да се екстраполира при възрастните – при тях растежът на височина е завършил, а теглото може да се увеличава (понякога и безконтролно).

Асоциация

Тестът за оценка на връзка между две качествени променливи се нарича χ^2 **тест за асоциация**. Той е един от най-често използваните в медицинската статистика. Тестът протича в следната последователност от стъпки:

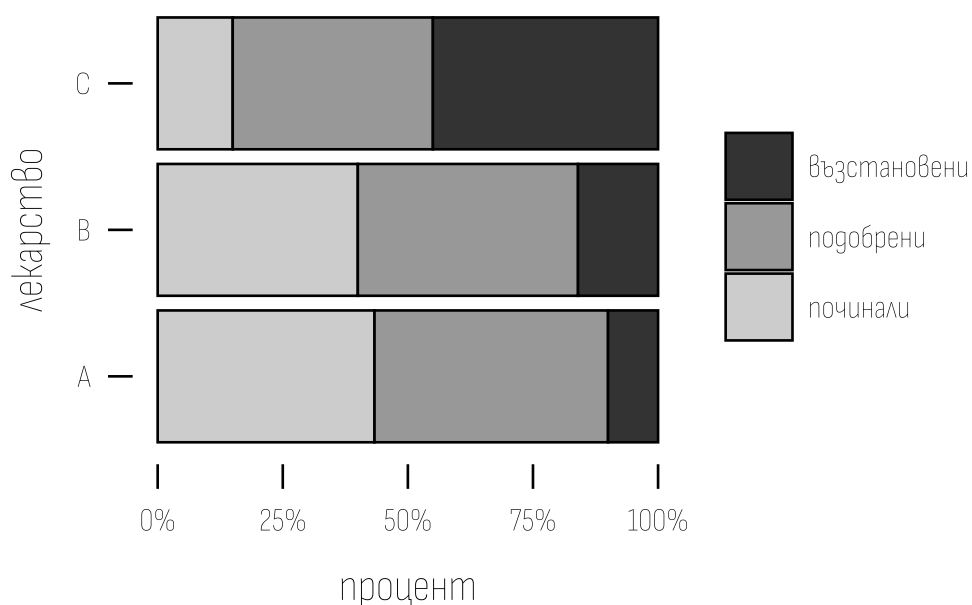
1. Дефиниране на нулева и алтернативна хипотеза;
2. Определяне на ниво на грешка;
3. Изчисляване на теоретични честоти;
4. Изчисляване на стойността на χ^2 ;
5. Определяне на р-стойност;
6. Заключение.

За прилагането му е представен следния пример:

Да приемем, че в проучване се тестват три вида лекарства за лечение на КОВИД-19. Медикаментите са означени като "А", "В" и "С". В изследването са включени 500 пациенти, от които лекувани с "А" са 150, с "В" – 250, а с "С" – 100. След 1 месец са установени следните резултати от терапията – починали са 180, с частично подобрение – 220, а напълно оздравели са 100. Резултатите са представени в таблица 1 и на фигура 1.

таблица 1: Данни (наблюдавани честоти) за лечението на COVID-19 с три лекарства при 500 изследвани лица

л-вство	починали	подобрене	оздравели	$\sum_{rows}$
A	f_{oa} 65	f_{od} 70	f_{og} 15	150
B	f_{ob} 100	f_{oe} 110	f_{oh} 40	250
C	f_{oc} 15	f_{of} 40	f_{oi} 45	100
$\sum_{columns}$	180	220	100	500



фигура 1: Наблюдавани честоти при изследване с три лекарства на 500 изследвани лица

Числата, записани във всяка една от клетките на таблица 1 (a, b, c и т.н) се наричат **наблюдавани** или **обсервирани** честоти. Те се бележат със символа f_o и са в резултат от действително наблюдаваните резултати от клиничното проучване.

Дефиниране на хипотези

Нулевата хипотеза твърди, че **не** съществува асоциация (връзка) между вида на терапията и резултата от лечението⁵.

Алтернативната хипотеза е противоположна на нулевата и твърди, че съществува асоциация между лекарството и резултата от лечението.

⁵ С други думи, каквото и хапче да даваме на тези пациенти, резултатът е един и същ, а ако наблюдаваме някаква разлика, тя е следствие на случайност.

Ниво на грешка

За ниво на грешка в повечето медицински проучвания се използва стойността 0,05. Това означава, че при бихме отхвърлили нулевата хипотеза, само ако вероятността тя да е вярна (да се наблюдавали връзката случайно) е по-малка от 0,05.

Изчисляване на теоретични честоти

Теоретичните честоти са числата (стойностите) на клетките (a, b, c и т.н), които **бихме наблюдавали**, ако нулевата хипотеза е вярна. Те се изчисляват по следната формула:

$$f_{tcell} = \frac{\sum_{columns} \cdot \sum_{rows}}{\sum_{total}}$$

Където:

- f_{tcell} е теоретична честота за всяка една клетка;
- $\sum_{columns}$ е сумата по колоните;
- $\sum_{rows}$ е сумата по редовете;
- $\sum_{total}$ е общата сума (всички участници);

Нека приложим формулата за първата клетка "a":

$$f_{ta} = \frac{180 \cdot 150}{500} = 54$$

Теоретичната честота f_{ta} означава, че ако липсва връзка между лечението и резултата, починалите пациенти лекувани с терапия "А" трябва да бъдат 54. Резултатите за всички клетки са представени в таблица 2 и на фигура 2.

таблица 2: **Теоретични** честоти за лечението на COVID-19 с три лекарства при 500 изследвани лица

л-вство	починали	подобрене	оздравели	$\sum_{rows}$
A	f_{ta} 54	f_{td} 66	f_{tg} 30	150
B	f_{tb} 90	f_{te} 110	f_{th} 50	250
C	f_{tc} 36	f_{tf} 44	f_{ti} 20	100
$\sum_{columns}$	180	220	100	500

За всички останали клетки

$$f_{tb} = \frac{180 \cdot 250}{500} = 90$$

$$f_{tc} = \frac{180 \cdot 100}{500} = 36$$

$$f_{td} = \frac{220 \cdot 150}{500} = 66$$

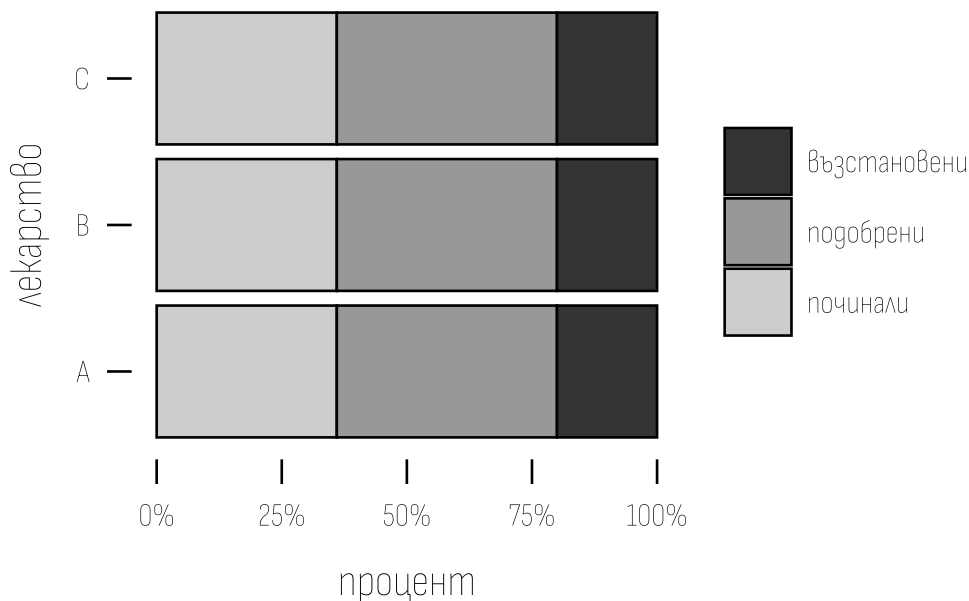
$$f_{te} = \frac{220 \cdot 250}{500} = 110$$

$$f_{tf} = \frac{220 \cdot 100}{500} = 44$$

$$f_{tg} = \frac{100 \cdot 150}{500} = 30$$

$$f_{th} = \frac{100 \cdot 250}{500} = 50$$

$$f_{ti} = \frac{100 \cdot 100}{500} = 20$$



фигура 2: Теоретични честоти, при вярна нулева хипотеза

Изчисляване на стойността на χ^2

Очевидно има разлика между очакванията на нулевата хипотеза (таблица 2) и реалните данни (таблица 1). За да се измери до каква степен наблюдаваните данни се различават от теоретичните честоти и да се оцени силата на връзка между изследваните променливи (вид лечение и клиничен резултат) се използва показателя χ^2 (хи-квадрат).

Формулата е:

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

В примера:

$$\chi^2 = \frac{(65-54)^2}{54} + \frac{(100-90)^2}{90} + \frac{(15-36)^2}{36} + \frac{(70-66)^2}{66} + \frac{(110-110)^2}{110} + \frac{(40-44)^2}{44} + \frac{(15-30)^2}{30} + \frac{(40-50)^2}{50} + \frac{(45-20)^2}{20}$$

$$\chi^2 = \frac{11^2}{54} + \frac{10^2}{90} + \frac{(-21)^2}{36} + \frac{4^2}{66} + \frac{0^2}{110} + \frac{(-4)^2}{44} + \frac{(-15)^2}{30} + \frac{(-10)^2}{50} + \frac{25^2}{20}$$

Във формулата:

- χ^2 е стойността на хи-квадрат;
- f_0 е наблюдаваната честота за всяка една клетка (a, b, c и т.н);
- f_t е теоретичната честота за всяка една клетка (a, b, c и т.н);
- $f_0 - f_t$ е разликата между наблюдаваната и теоретичната честота за всяка една клетка (a, b, c и т.н);
- $(f_0 - f_t)^2$ е разликата повдигната на квадрат;
- $\frac{(f_0 - f_t)^2}{f_t}$ е делимото на повдигната на квадрат разлика спрямо теоретичната честота за всяка една клетка (a, b, c и т.н);
- $\sum$ е сумата от всички разделяния.

$$\chi^2 = \frac{121}{54} + \frac{100}{90} + \frac{441}{36} + \frac{16}{66} + \frac{0}{110} + \frac{16}{44} + \frac{225}{30} + \frac{100}{50} + \frac{625}{20}$$

$$\chi^2 = 2,24 + 1,11 + 12,25 + 0,24 + 0 + 0,36 + 7,5 + 2 + 31,25 = 56,95$$

Стойността **54** е оценка на "несъответствието" между това, което нулевата хипотеза твърди и това, което ние действително сме наблюдавали.

Определяне на р-стойност

Следва да се определи, каква е вероятността това несъответствие да е чиста случайност p . По подобие на t стойността в упражнение 3, отново се използва таблица, с известни стойности χ^2 и съответните им вероятности p . Преди тази стъпка трябва да определяме "степените на свобода" df . От тях зависи точния ред на статистическата таблица за χ^2 разпределението, който трябва да се използва за всеки конкретен пример.

df е число, което зависи от размера на таблично представените данни - броя на колоните и редовете и се изчисляват по формулата:

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

В посочения пример - таблицата е с 3 колони и 3 реда (таблица 3x3).

Степените на свобода са:

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

В таблица 3 са представени данните само за реда, съответстващ на $df = 4$. За всяка стойност на χ^2 е определена вероятността p .

В тази стъпка сравняваме изчислената стойност на $\chi^2 = 57$ с наличните данни. Въпреки че точната стойност не е налична в таблица 3, тя е по-голяма от **най-дясно разположеното** число - 13,27, което съответства на вероятност p 0,01.

Във формулата:

- $N_{columns}$ е броят на колоните;
- N_{rows} е броят на редовете;

таблица 3: Таблица за определяне на р-стойността при χ^2 тест за асоциация

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



Важно

С увеличаването на стойността на χ^2 , вероятността p намалява.

Заклучение

$\chi^2 = 57$ е по-голямо от 13,27, което означава, че вероятността p е по-малка от нивото на грешка 0,05. Следователно – отхвърляме нулевата хипотеза и приемаме алтернативната. Има статистически значима **връзка (зависимост)** между избраното лечение и клиничния резултат.

Малко повече за асоциацията

- За да прилагаме този тест е необходимо да имаме поне **5** наблюдения във всяка една клетка. При по-малки стойности не можем да се доверим на получената стойност за p .
- Когато работим с качествени (алтернативни) признаци с две възможни стойности (таблицы 2x2) може да се използва разновидност на *корелационния анализ*. Коефициентът в случая се означава с символът ϕ ⁶, а пример за неговото приложение е представен в таблица 4.

⁶ ϕ се интерпретира, както коефициентът на Пирсън – обяснен по-надолу.

таблица 4: Таблица 2x2 за изследване на връзката между ваксината и имунния отговор

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

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

В примера:

$$\phi = \frac{(90 \cdot 90) - (160 \cdot 10)}{\sqrt{(90 + 160) \cdot (10 + 90) \cdot (90 + 10) \cdot (160 + 90)}}$$

$$\phi = \frac{(8100) - (1600)}{\sqrt{(250) \cdot (100) \cdot (100) \cdot (250)}}$$

$$\phi = \frac{6500}{\sqrt{62500000}} = \frac{6500}{7906,25} = 0,82$$

Където:

- ϕ е коефициентът на корелация;
- "a" е броят на ваксинираните с новата ваксина и придобили имунитет;
- "b" е броят на ваксинираните с старата ваксина и и придобили имунитет;
- "c" е броят на ваксинираните с новата ваксина и не придобили имунитет;
- "d" е броят на ваксинираните със старата ваксина и не придобили имунитет;

Корелация



Корелационният анализ е:

Статистически метод, използван за изследване на взаимовръзка между две **количествени** променливи

Корелационен коефициент

Корелационният коефициент е число в диапазона от -1 до +1. Колкото по отдалечена е стойността му от 0 –тата, толкова **по-силна** е наблюдаваната статистическа връзка. Съществуват множество методики и видове корелационни коефициенти, въпреки това целта на всички тях е:

- Да оцени **силата на статистическата връзка**
- Да се оцени **посоката** на връзката⁷;
- Да се определи **формата** на връзката⁸;

Коефициент на Пирсън

В настоящият курс е представен само един от множеството коефициенти за корелация – този на Пирсън, който се означава с гръцката буква ρ или с латинската r . Коефициентът е подходящ, когато изследваните променливи са **нормално разпределени**, а предполагаемата връзка между тях е **линейна**. Интерпретацията на коефициента на Пирсън е представена в таблица 5.

⁷ Права връзка (позитивна корелация) се наблюдава, когато двете променливи се увеличават или намалят синхронно. Негативна корелация (обратна връзка) се наблюдава, когато увеличението в една променлива е свързано с намаление в другата.

⁸ Връзката може да е “линейна” още пропорционална или “нелинейна” – експоненциална, логаритмична и т.н. В този курс ще се спрем единствено на **линейната връзка**.

Изчисление на коефициента на Пирсън

ρ се изчислява с помощта на следната формула:

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

Където:

- x_i и y_i са стойностите на променливите x и y за всеки един от участниците в изследването;
- $\bar{x}$ и $\bar{y}$ са средните аритметични за променливите x и y ;

таблица 5: ρ интерпретация

ρ	сила и посока на връзката
1	“перфектна” <i>права</i>
0,67 до 1	силна <i>права</i>
0,34 до 0,66	умерена <i>права</i>
0 до 0,33	слаба <i>права</i>
0	липсва връзка
0 до -0,33	слаба <i>обратна</i>
-0,34 до -0,66	умерена <i>обратна</i>
-0,67 до -1	силна <i>обратна</i>
-1	“перфектна” <i>обратна</i>

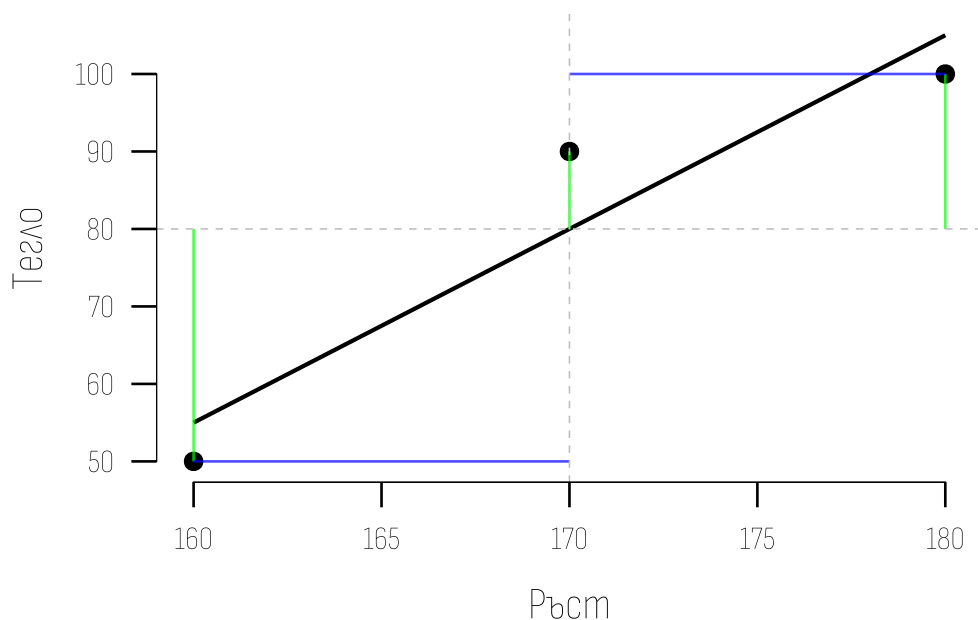
- $x_i - \bar{x}$ е отклонението на всяка една точка от средната аритметична за променливата x ;
- $y_i - \bar{y}$ е отклонението на всяка една точка от средната аритметична за променливата y ;
- $\sum_{i=1}^n (x_i - \bar{x})^2$ е сумата на квадратите на отклоненията за променливата x ;
- $\sum_{i=1}^n (y_i - \bar{y})^2$ е сумата на квадратите на отклоненията за променливата y ;

Пример за изчисление на коефициента на Пирсън

Формулата може да се представи като последователност от стъпки. Нека приемем, че разполагаме с данните за ръста и теглото на 3 студенти и следва да определим корелацията между тези две количествени променливи. Данните са представени в таблица 6 и на фигура 3.

таблица 6: Данни за ръста и теглото на 3-ма студенти

име	Ръст (x)	Тегло (y)
Мария	160	50
Иван	170	90
Чавдар	180	100



фигура 3: Диаграма на разсейване за ръста и теглото на 3 участника в изследване. Със сиви прекъснати линии – средните аритметични, с черна линия – корелационната връзка, с сиви (зелени и сини) непрекъснати линии – отклоненията

Стъпка 1 – изчисляване на средната аритметична

Изчисляваме средната аритметична, както за променливата x (ръста), така и за променливата y (теглото).

за ръста

$$\bar{x} = \frac{160 + 170 + 180}{3} = 170$$

за теглото

$$\bar{y} = \frac{50 + 90 + 100}{3} = 80$$

Стъпка 2 - Изчисляване на отклоненията

Отклоненията представляват разликата между стойността на наблюденията и средната аритметична за групата.

За първия студент (Мария) тези отклонения са:

за ръста

$$x_{i=1} - \bar{x} = 160 - 170 = -10$$

за теглото

$$y_{i=1} - \bar{y} = 50 - 80 = -30$$

В таблица 7 са представени резултатите от тази стъпка за всички наблюдавани студенти.

таблица 7: Таблица с отклоненията на ръста и теглото на 3-ма студенти

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

За втория студент (Иван) тези отклонения са:

$$x_{i=2} - \bar{x} = 170 - 170 = 0$$

$$y_{i=2} - \bar{y} = 90 - 80 = 10$$

За третия студент (Чавдар) тези отклонения са:

$$x_{i=3} - \bar{x} = 180 - 170 = 10$$

$$y_{i=3} - \bar{y} = 100 - 80 = 20$$

Стъпка 3 - Изчисляване на кръстосания продукт

В трета стъпка се определя "сумата на кръстосания продукт." Под "кръстосан продукт" се разбира умножението на отклоненията за всяка една от променливите.

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

За първия студент (Мария) кръстосаният продукт е:

$$(x - \bar{x}) \cdot (y - \bar{y}) = (-10) \cdot (-30) = 300$$

таблица 8 представя таблично резултатите от тази стъпка и сумата на кръстосаните произведения за всички наблюдавани студенти.

таблица 8: Таблица с кръстосани произведения на ръста и теглото на 3-ма студенти

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
	$\bar{x} = 170$	$\bar{y} = 80$			$\sum = 500$

Стъпка 3 предоставя числителя на посочената по-горе формула. В използвания пример той е равен на 500.

Стъпка 4 - Изчисляване на сумата на квадратите на отклоненията

Стъпка 4 се използва за определяне на знаменателя на посочената формула. В нея, изчисляваме сумата на квадратите на отклоненията за всяка една променлива - $(x - \bar{x})^2$ и $(y - \bar{y})^2$.

Самите отклонения са изчислени в стъпка 2 и представени в таблица 7. В този етап, те се повдигат на квадрат и сумират, както е представено в таблица 9.

таблица 9: Таблица с квадратите на отклоненията на ръста и теглото на 3-ма студенти

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
	$\bar{x} = 170$	$\bar{y} = 80$		$\sum = 200$		$\sum = 1400$

За втория студент (Иван):

$$(x - \bar{x}) \cdot (y - \bar{y}) = (0) \cdot (10) = 0$$

За третия студент (Чавдар):

$$(x - \bar{x}) \cdot (y - \bar{y}) =$$

$$= (10) \cdot (20) = 200$$

Стъпка 5 – Изчисляване на корелационния коефициент

Вече сме готови за заместим във формулата:

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

$$\rho = \frac{500}{\sqrt{200} \cdot \sqrt{1400}}$$

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

Стъпка 6 Заключение

При тези трима студенти се наблюдава **силна позитивна** корелация между ръста и теглото – увеличението на теглото се свързва с увеличение на ръста. Връзката е валидна **само за тези трима студенти**. За да установим дали има такава в генералната съвкупност, провеждаме тест на хипотеза⁹

В случая:

Нулевата хипотеза твърди, че популационният корелационен коефициент е **нула**. С други думи не съществува корелация между теглото и ръста.

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

За да тестваме нулевата хипотеза, използваме тест подобен на вече известния Т теста.

⁹ Тестът на хипотеза за корелационния коефициент не е обект на настоящия курс.