01-Упражнение 2023/2024

Основна терминология. Стандартизация

ас.д-р Костадин Костадинов

Защо учим статистика?

- Статистиката е метод чрез който данните се превръщат в информация.
- Статистиката помага в ежедневната медицинска практика, чрез създаване на клинични наръчници или политики в здравеопазването.
- Статистиката е езикът на науката, тя помага както на пациентите, така и за поддържане на здравето в обществото.

С какво ще ни помогне статистиката?

- Да вземаме **информирани решения** в ежедневната ни практика.
- Да разберем как работи **науката**.
- За да четем критично **нова** научна информация.

Какво няма да научите?

Понастоящем статистиката е основата на в т.н "наука за данните" (data science). В съчетание със сложна математика, програмиране и доза креативност, тази нова дисциплина навлиза във всекидневния ни живот докато пазаруваме, учим и работим. Всички технологии базирани на "изкуствен интелект" същност използват статистика [1] . В този курс, обаче целта е да придобиете най-основните знания за това как работи статистиката, каква е логиката в нея и какъв език използва.

[1] В момента дори има разработени скенери които "сами разчитат" дали пациента има заболяване и показват какво е то. Разработени са и електрокардиографии, записващи сърдечната дейност на пациента и "автоматично" разпознаване дали е налице определено заболяване.

Терминология

За да "не сме изгубени в превода", въвеждаме някои основни термини, обяснени с примери.

Абсолютни величини

> ⚠ Определение
>
> Това са **числа**, които количествено характеризират обемите на статистическите съвкупности или на части от тях. Те представляват стойност на конкретни статистически признаци.

- Абсолютните величини са винаги **наименовани** с конкретни **мерни единици**.
- Статистическите изследвания, обикновено започват с анализ на абсолютните величини, но те **не са достатъчни** за директни сравнения в **пространствено-времеви аспект.**

Примери за абсолютни величини

Систолното артериално налягане измерено в *mmHg* е абсолютна величина - има числова стойност, мерна единица и количествено характеризира определен признак. Кръвната захар измерена в mmol/l също е абсолютна величина - отново е число с мерна единица и измерващо конкретен показател.

Относителни величини

> ⚠ Определение
>
> Те се изчисляват при разделяне на две абсолютни величини. Представят се като коефициенти, а при умножение по 100 или 1000 в проценти или промили.

Примери за относителни величини

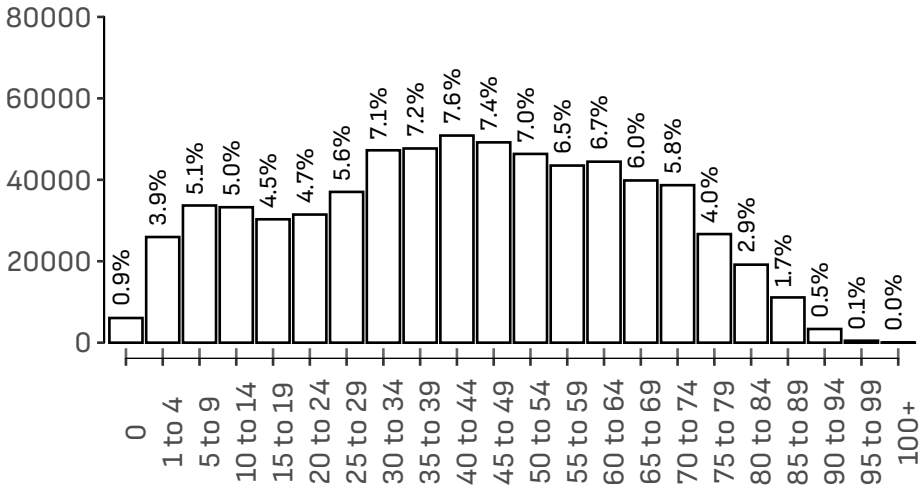
В медицината, често използваме относителни величини. Например, когато измерваме помпената функция на сърцето можем да измерим количеството кръв, което постъпва в аортата, след едно сърдечно съкращение. Това е абсолютната величина **ударен обем**. Хората с по-висок ръст и по-високо тегло (по-едро телосложение) имат по-високи стойности на усърдния обем, спрямо тези с по-нисък ръст и по-малко тегло. Така например сърцето на състезател по сумо изтласква по-голямо количество кръв (в милилитри), спрямо сърцето на първокласник. Означава ли това, че сърцето на сумиста работи по-добре от това на първокласника? Отговорът е, че не можем да преценим - двете абсолютни величини не бива да се сравняват директно. Затова по-важното, в случая е съотношението на ударният обем, спрямо количеството кръв налично в сърцето, точно преди неговото съкращение. Това е т.н "фракция на изтласкване" и представлява **относителна величина**.

Екстензивни показатели

> Определение
>
> Това са **структурни** показатели, които показват как едно статистическо явление се разпределя на съставните си части в определено време и място.

Примери за екстензивни показатели

Ако приемем "възрастта" в гр. Пловдив за статистическо "явление" можем да представим всички жители на града в категории по възрастова група - новородени до 1 г., деца между 1 и 5 год., от 5 до 10 г. и т.н. Ако разделим броя на хората в съответната възрастовата група, спрямо всички жители на града ще получим екстензивен показател - измерен в процент. На фигура 1 е представено разпределението на възрастта в гр. Пловдив. **Важно** за екстензивните показатели е, че сумата от всички тях е равна на 1 (или 100%).



фигура 1: Пример за екстензивен показател.Възраст на населението в гр. Пловдив, 2022 г.

Интензивни показатели

⚠ Определение

Това са **честотни** показатели, които показват колко често се среща дадено явление в свойствената си среда. Всеки интензивен показател е съотношение между обемите на две **различни** статистически съвкупности, намиращи се във връзка една с друга. В числителя е явлението от което се интересуваме, а в знаменателя е абсолютният обем на средата, в която възниква то.

Примери за интензивни показатели

1. **Леталитетът** представлява броя смъртни случаи от конкретно заболяване, спрямо общия брой болни от това заболяване. Ако леталитетът от морбили (дребна шарка) при деца (до 18 г.) е 5%, това означава, че от 100 деца със заболяването, 5 са починали. С други думи, показателят представя **честотата** на една статистическа съвкупност (смъртните случаи) върху друга съвкупност (болните деца).
2. **Смъртността** представлява съотношение на броя на починалите, спрямо средния брой население. **Смъртността по причини** представлява броя на починалите от дадено заболяването, разделен на броя на всички починали. Двата показателя не бива да се бъркат с леталитета.

3. **Заболеваемостта**, представлява съотношението на броя новозаболели от дадено заболяване (например от рак на гърдата), спрямо популацията в риск (всички, които биха могли да се разболеят от това заболяване).

4. В ежедневната практика като лекари, също ще ползвате интензивни показатели. Например, при пациенти с белодробна астма, видът на използваното лечение, зависи от честотата на екзацербации (обостряния) т.н интензивния показател "exacerbation rate".

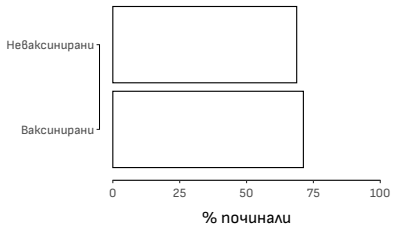
Пряк метод на стандартизация

Преди определението за стандартизация, нека представя следния пример пример. Представете си, че от днес сте министър на здравеопазването. Изправени сте пред сериозен проблем - в държавата върлува опасен вирус. Имате ваксина, но липсва доверие на гражданите в нея. Много хора смятат, че ваксините дори убиват. Днес след среща с граждани, противопоставящи се на ваксините, получавате научна статия, в която се твърди, че ваксините повишават вероятността от смърт. В таблица 1 можете да видите данните от нея.

таблица 1: Таблица на антиваксарите

	Общо	Починали
Ваксинирани	4000	2850
Неваксинирани	8000	5500

Авторът посочва, че след 1 година от 4000 ваксинирани са починали 2850 души, докато при 8000 неваксинирани, починали са 5500. Тези числа представляват абсолютни величини. За да се сравнят, трябва да използваме относителни. С други думи, каква пропорция от ваксинираните са починали, спрямо тази при неваксинираните. "Сметката" тук е лесна: трябва да разделим броя на починалите сред ваксинираните, върху общият брой на ваксинираните. Същата пропорция трябва да изчислим и за неваксинираните. Резултатите са относителни величини - интензивни показатели.



фигура 2: Разлика между смъртността при ваксинирани и неваксинирани.

Оказва се, че в групата на ваксинираните починалите са 71.2%, докато при неваксинираните - 68.8%. Това е разлика от 2,4 процентни пункта[2]. Може би, наистина "антиваксърите" имат право.

Как бихме могли да си обясним този резултат? Нима наистина ваксините са причина за по-големия брой смъртни случай? Трябва ли да продължим да използваме тази ваксина? Бихте ли посъветвали пациентите си да се ваксинират?

Преди да дадем категоричното си решение, трябва да помислим върху данните. Те все още не са информация, на която да базираме решенията си. В случая, можем да разглеждаме цифрите в таблицата, като *сурови* данни измерващи една връзка. Това е връзката между ваксинация и смъртен изход. Не изглежда логично, смъртта да се причинява единствено от ваксината или липсата на такава. Има редица други фактори, които могат да повлияят - например **възрастта**.

[2] Важно: простите аритметични операции между проценти се изразяват в процентни пунктове.

Нормално е, ако ваксинираните са по-възрастни спрямо неваксинираните, при тях да наблюдаваме повече починали, дори и ако ваксината наистина работи и ефективна.

При сравняването на интензивни величини се наблюдава фактът, че стойността на тези показатели е в зависимост от структурата на средата, в която са изучавани явленията. За да проверим дали тази среда "замъглява" връзката между фактора и резултата, можем да използваме статистическия метод на **стандартизацията**.

> ⚠ Определение
>
> Под стандартизация се разбира преобразуването на общите коефициенти, с което се отстранява (елиминира) влиянието на възрастови или други различия в състава на сравняваните групи.

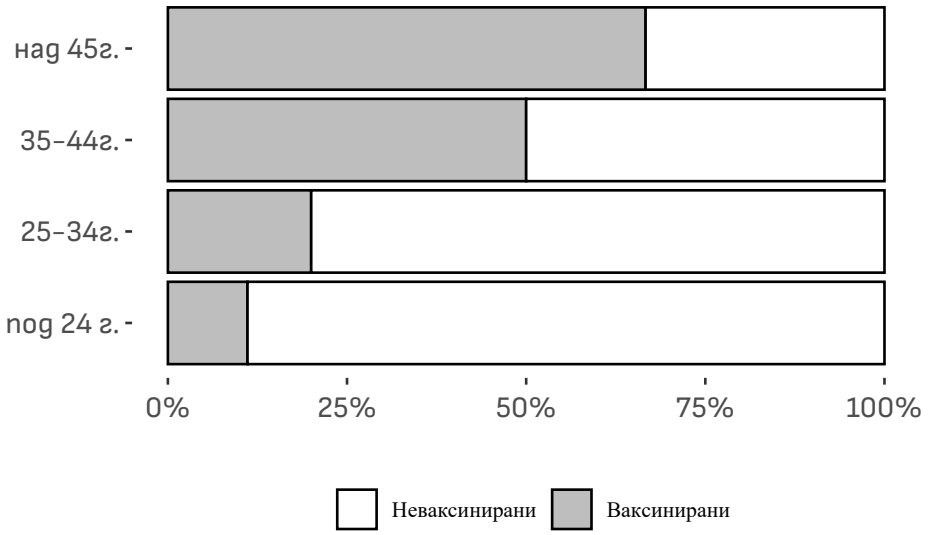
Стъпки

За да извършим стандартизация (в курса по статистика се спираме единствено на **прекия метод за стандартизация**) следва да разполага с повече данни. Таблицата, която разгледахме, не съдържа информация за възрастта на участниците. Затова, след запитване към автора, получаваме по-подробни данни - представени в таблица 2 .

таблица 2: Таблица с разпределние по възраст

	Възраст	Общо	Починали
Ваксинирани	под 24 г.	500	250
	25-34г.	500	300
	35-44г.	1000	700
	над 45г.	2000	1600
Неваксинирани	под 24 г.	4000	2400
	25-34г.	2000	1400
	35-44г.	1000	800
	над 45г.	1000	900

Може би, ви прави впечатление от фигура 3, че ваксинираните, са предимно по-възрастни хора, докато при неваксинираните преобладават по-младите.



фигура 3: Разлика на възрастовата структура между ваксинирани и неваксинирани - възрастовите групи са представени от млади към стари.в групата до 24г. преобладат неваксинирани, докато при над 45-годинишните ваксинираните.

Стъпка 1 Изчисляване на нестандартизираните интензивни показатели

Както по-рано, така и сега, можем да изчислим, какъв процент от участниците в двете групи са починали. В случая ще направим това за всяка една възрастова група. Резултатите са представени в колона *леталитет* в таблица 3 .

	Възраст	Общо	Починали	Леталитет[1]
Ваксинирани	под 24 г.	500	250	0.5
	25-34г.	500	300	0.6
	35-44г.	1000	700	0.7
	над 45г.	2000	1600	0.8
Неваксинирани	под 24 г.	4000	2400	0.6
	25-34г.	2000	1400	0.7
	35-44г.	1000	800	0.8
	над 45г.	1000	900	0.9

таблица 3: Нестандартизиран леталитет. Тук е важно да запомним, че общият нестандартизиран показател не е сума от резултатите по групи. Не можем да сумираме нестандартизираните показатели.

[1]Нестандартизиран

Стъпка 2 Изчисляване на "стандарт"

За да направим стандартизацията е необходимо да изберем за стандарт, една от двете възрастови структури - тази на ваксинираните или тези на неваксинираните.

Тук, често възниква въпросът коя структура трябва да изберем? Защо да предпочетем едната спрямо другата? Какво е правилото?

Всъщност, няма особено значение точно коя структура се избира. Разбира се, числата след стандартизация зависят от избора и те биха се различавали. От значение, обаче е разликата, а не конкретните стойности на стандартизиран леталитет в двете групи [3]

[3] В това упражнение ще докажем това, като извършим стандартизацията, като вземем за стандарт, първо възрастовата структура на неваксинираните, а после тази на ваксинираните.

Какво обаче е стандартът?

Ако изберем за стандарт групата на неваксинираните, за да изчислим стандарта - ще използваме броя на участниците във всяка възрастова група за числител, а общия брой неваксинирани за знаменател. Полученият коефициент е "стандарт" за конкретната възрастова група.

В таблица 4 са представени получените стандарти.

Неваксинираните участниците под 24 год. са 4000, а общият брой неваксинирани 8000. Стандартът за тази група е 0.5 (ако умножим по 100 ще получим 50%). Това е стандартът за тази група, които обаче ще използваме и за ваксинираните.

Може да ви направи впечатление, че сборът на всички стандарти е равен на 1-ца (тоест 100%). Това е така, защото стандартът винаги е екстензивен показател.

таблица 4: Определяне на стандарт за всяка възрастова група

Възраст	Общо	Починали	Стандарт[1]
под 24 г.	4000	2400	0.500
25-34г.	2000	1400	0.250
35-44г.	1000	800	0.125
над 45г.	1000	900	0.125

[1]За изчисляване на колоната стандарт е изпозлвана възрастовата структура на неваксинираните

Стъпка 3: Изчисляване на стандартизираните показатели

За да изчислим стандартизираният леталитет, за всяка възрастова група **умножаваме** нестандартизирания показател по посоченият по-горе стандарт.

На този етап стандартизацията е почти изпълнена - Вече знаем, че **нестандартизираните показатели не се събират**. За сметка на това **стандартизираните се събират**. Когато ги съберем получаваме общия стандартизиран леталитет, който е "изчистен" от замъгляващия ефект

таблица 5: Например, за възрасттовата група до 24 г. при ваксинираните, нестандартизирания леталитет е 0,5, а стандартът 0,250. За да стандартизираме умножаваме 0,5 по 0.250. При неваксинираните, отново за същата възрастова група умножаваме нестандартизирания показател 0,6 по стандарта 0,25

	Възраст	Общо	Починали	НС Леталитет[1]	Стандарт[2]	С Леталитет[3]
Ваксинирани	под 24 г.	500	250	0.5	0.500	0.2500
	25-34г.	500	300	0.6	0.250	0.1500
	35-44г.	1000	700	0.7	0.125	0.0875
	над 45г.	2000	1600	0.8	0.125	0.1000
Неваксинирани	под 24 г.	4000	2400	0.6	0.500	0.3000
	25-34г.	2000	1400	0.7	0.250	0.1750
	35-44г.	1000	800	0.8	0.125	0.1000
	над 45г.	1000	900	0.9	0.125	0.1125

[1]Нестандартизиран леталитет
[2]Изчислен спрямо неваксинираните
[3]Стандартизиран леталитет

на различната възраст в двете групи. С други думи, показателите след стандартизация, представят какъв би бил леталитетът, ако двете групи имаха еднаква възрастова структура.

Стъпка 4: Заключение

Нека извършим тази последна калкулация.

За групата на ваксинираните общият стандартизиран леталитет е:

- **25 %** (стандартизираният леталитет за всички до 24 г.) + **15 %** (стандартизираният леталитет за възрастовата група от 25-34 г.) + **8,7 %** (стандартизираният леталитет за възрастовата група от 35-44 г.) + **10 %** (стандартизираният леталитет за възрастовата група над 45 г.). Общо за всички ваксинирани, стандартизираният леталитет е 58.7%

За групата на неваксинираните общият стандартизиран леталитет е:

- **30 %** (стандартизираният леталитет за всички до 24 г.) + **17,5 %** (стандартизираният леталитет за възрастовата група от 25-34 г.) + **10 %** (стандартизираният леталитет за възрастовата група от 35-44 г.) + **11,25 %** (стандартизираният леталитет за възрастовата група над 45 г.). Общо за всички ваксинирани, стандартизираният леталитет е 65.75%

Заключение

Стандартизираните показатели за леталитет в двете групи са съответно **58.75 %** и **68.75 %**. Леталитетът сред неваксинираните, е с **10 %** по-висок.

Стандартизираните показатели позволяват да се анализира и оцени нивото на изучаваното явление при създадени условия на **еднородност**, тоест методът ни показва какви биха били коефициентите, в сравняваните групи, ако те имаха еднакъв състав.

Стъпки - при алтернативен избор за стандарт

За да докажем, че изводът не зависи от избора на стандарт, ще решим отново примера, като този път използваме за стандарт възрастовата структура на ваксинираните.

Стъпка 2 Изчисляваме стандарта (този път спрямо ваксинираните)

Сега ще използваме данните само за ваксинираните. При тях, частниците под 24 г. са 500 от общо 4000. Това означава, че стандартът за тази група е 0.125 (или 12.5%). Получените стандарти са представени в таблица 6 .

таблица 6: Определяне на стандарт за всяка възрастова група

Възраст	Общо	Починали	Стандарт[1]
под 24 г.	500	250	0.125
25-34г.	500	300	0.125
35-44г.	1000	700	0.250
над 45г.	2000	1600	0.500

[1]За изчисляване на колоната стандарт е използвана възрастовата структура на ваксинираните

Стъпка 3: Изчисляване на стандартизираните показатели за леталитет

След като имаме "стандарт", този път в основа на групата на ваксинираните, можем да пристъпим отново към стъпка 3 - стандартизация тя е представена в таблица 7

таблица 7: Отново, за възрасттовата група до 24 г.при ваксинираните, нестандартизирания леталитет е 0,5, а новият стандарт 0,125. Стандартизираният леталитет е 0,5x0,125 = 0,0625. При неваксинираните, до 24 г. нестандартизирания леталитет е 0,6, за да получим стандартизирания умножаваме по стандарта 0,125, получаваме 0,075)

	Възраст	Общо	Починали	НС Леталитет[1]	Стандарт[2]	С Леталитет[3]
Ваксинирани	под 24 г.	500	250	0.5	0.125	0.0625
	25-34г.	500	300	0.6	0.125	0.0750
	35-44г.	1000	700	0.7	0.250	0.1750
	над 45г.	2000	1600	0.8	0.500	0.4000
Неваксинирани	под 24 г.	4000	2400	0.6	0.125	0.0750
	25-34г.	2000	1400	0.7	0.125	0.0875
	35-44г.	1000	800	0.8	0.250	0.2000
	над 45г.	1000	900	0.9	0.500	0.4500

[1]Нестандартизиран леталитет
[2]Изчислен спрямо неваксинираните
[3]Стандартизиран леталитет

Логично, след като сме използвали друг стандарт, числовите стойности са различни, но заключението едно и също.

Заключение

Стандартизираните показатели за леталитет в двете групи са съответно **71.25 %** и **81.25 %**. Леталитетът сред неваксинираните, е с **10 %** по-висок.

01-Упражнение 2023/2024

Домашна работа

ас.д-р Костадин Костадинов

Домашна работа

1. Да се определи възрастовата структура на преминалите болни през Болница Х, като разполагате със следните данни:

Възраст	0-1 г.	1-4 г.	5-9 г.	10-19 г.	Общо
Болни с Хепатит А	70	41	102	87	300
Възрастова структура (%)					

2. Да се изчислят стандартизираните показатели за плодовитостта в районите А и Б. За стандарт да се приеме възрастовият състав на жените от район Б. Да се анализират получените резултати.

	Район А		Район Б	
Възраст	Жени	Живородени	Жени	Живородени
15-19	1 000	18	1 200	22
21-30	9 000	225	7 000	175
31-49	8 000	128	10 000	160
Всичко	18 000	371	18200	357

3. Да се изчислят стандартизираните показатели за леталитета в две градски болници - Болница А и Болница Б. За стандарт да се приеме съставът на болните в болница Б. Да се анализират получените резултати.

Болест	Преминали (А)	Починали (А)	Преминали (Б)	Починали (А)
Хипертонична болест	180	4	200	4
Рак на стомаха	100	30	90	27
Инфаркт на миокарда	120	8	160	10
Всичко	400	42	450	41

4. Нестандартизираните показатели за леталитет в две болници А и Б са съответно 10‰ и 12‰. След стандартизация, спрямо болница А, леталитетът в болница А е 10‰, а този в болница Б 8‰. Изчислете стандартизираните леталитети, ако за стандарт се ползва болница Б.

02-Упражнение 2023/2024

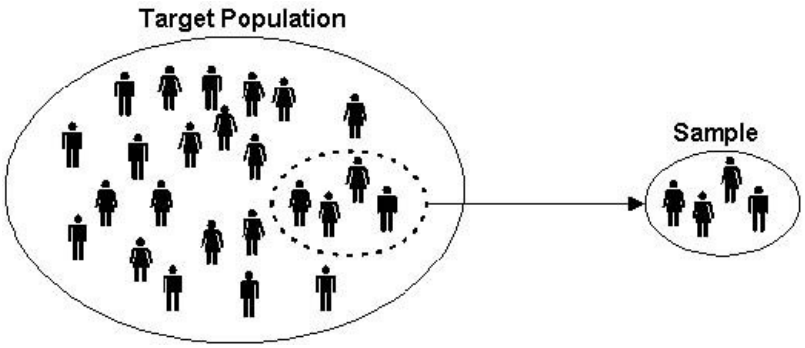
Дескриптивна и инферентна статистика

ас.д-р Костадин Костадинов

Увод

Провеждането на едно научно изследване е трудоемък процес, в който се изискват се множество финансови, човешки и времеви ресурси. Много често не е възможно изследването на всички индивиди които са обект на интерес. Немислимо е например да се прегледат всички пациенти със сърдечна недостатъчност, за да се установи дали ново лекарство намалява пулса им. В помощ на науката идва статистиката. Тя ни позволява изследваме малък брой единици (**наречени извадки**), а резултатите от тях да обобщаваме за всички единици в т.н. **генералната съвкупност** [1]. На фигура фигура 1 са представени схематично посочените термини.

[1] За да направим това "обобщение" е необходимо извадката да бъде представителна - тоест, да възпроизвежда структурата (най-малко по пол и възраст) на генералната съвкупност, а наблюдаваните единици да са избрани случайно.



фигура 1: Генерална съвкупност и извадка

Основни понятия

⚠ Дескриптивна и инферентна статистика

Дескриптивна (описателна) статистика представлява анализ и описание на статистическите явления **в извадка**, а инферентна статистика е процесът на генерализация (обобщение) на получените резултати в **генералната съвкупност**

Статистически променливи

Статистическите данни се "вземат" чрез измерване на определена характеристика върху група от обекти. Тъй като измерванията могат да приемат различни числови или нечислови стойности, те се наричат още **променливи**. За една група студенти от втори курс например, може да се установи, че се различават по пол, възраст, успех, отношение към определен проблем и редица други **характеристики**. Следователно изброените са наблюдавани **променливи**. От друга страна, ако характеристиката е една и съща за всеки член на групата, тя се нарича **константа**. В нашия пример цялата група се състои от студенти II курс, затова тази характеристика (курс на обучение) е константа.

⚠ Статистическият признак

Представлява отличителна черта, белег на една или повече единици - обекти на статистически изследвания.

Статистическите признаци се различават по вид, което налага различен подход при статистическото им изучаване. Основното им разделяне е на **вариационни** и **категорийни**. Вариационните, наричани още **количествени**, са тези, които се изразяват с **числа**. Те могат да бъдат **прекъснати (дискретни)** и **непрекъснати (континюитетни)**. Прекъснатите могат да приемат отделни, изолирани една от друга числови стойности. Такива са броят на членовете на отделно домакинство, броят боледувания за една година, брой живородени деца и др. [2]. Непрекъснати са признаците, които могат да получават всякакви характеристики в даден числов интервал. Такива са: теглото, температурата, кръвното налягане. Посочените характеристики могат да бъдат измервани с изключителна точност след десетичната запетая. Категорийните признаци, наричани още атрибутивни или качествени са тези,

[2] Въпреки, че прекъснатите характеристики не се изразяват с дробни числа, много често това се налага, когато те се анализират като прекъснати променливи. Така, въпреки че за някого звучи абсурдно, може да кажем, че средностатистическата българка ражда 1,7 деца.

чиито характеристики нямат числов израз. За тях се използват "етикети" или описателно наименование. Такива са например пол, семейно положение, кръвна група, вид заболяване и др.

⚠ Признаците биват два основни типа: *количествени и качествени*.

Количествени са тези променливи, които могат да бъдат изразени числено, а **качествени** са тези, които нямат числено значение.

Скали за измерване

Скалата, по която се измерва една статистическа променлива, е фактор от голямо значение при определянето на подходящи методи за анализ. Разделянето на скалите за измерване се извършва в съответствие със степента на точност, която използват [3]. Качествените променливи се измерват на **номинална, рангова и ординална** скала, а количествените - в **интервална и/или пропорционална** скала (скала на отношенията). Първите две скали се наричат неметрични и се определят като "слаби, неточни скали" а вторите две — метрични или "силни, точни скали". Описание на скалите с примери от медицинската статистика са представени в таблица 1

[3] Ако кажем за един човек, че е висок, това не е равнозначно да кажем, че неговият ръст е 185 см. В първия случаи използваме неточно измерване на "номинална скала", а във втория - точно на "пропорционална скала"

таблица 1: Видове признаци и скали на измерване

Скала	Характеристика	Пример
Номинална	Категории, "наименования, етикети".	Диагноза, кръвна група
Дихотомна	*Вид номинална скала за признаци само две възможни стойности.*	*пол*
Ординална	Качествени променливи, които имат логически (нарастващ или намаляващ) ред. Не е възможно изчисление на пропорция или разлика	Степен на тежест на сърдечна недостатъчност (лека, умерена, тежка), стадии на онкологично заболяване (I-ви, II-ри ..)

Скала	Характеристика	Пример
Рангова	*Вид ординална скала при която единиците се записват с рангове (поредни места) от 1 до n.*	*Оценки "слаб 2", "среден 3" и т.н.*
Интервална	Количествени променливи. Стойността "0" не означава липса на признака, тя е допустима наред с други позитивни и отрицателни стойности. Не позволява отношения	Температура (NB: Ако днес е 10 градуса по-топло от вчера, няма как да изчислим "колко процента по-топло е")
Пропорционална	Количествени променливи. Стойността "0" означава липса на признака. Позволява съотношения	Тегло (можем да го измерим с голяма точност след десетичната запетая).

Дескриптивен статистически анализ

Статистическият анализ обхваща проучване посредством подходящи методи на **обобщаващи числови характеристики** и смисловото им интерпретиране. Тези характеристики се разделят на два вида - описващи **централната тенденция** и **разсейването** в изследваната извадка [4]. Централната тенденция при променливите, които се измерват на **пропорционална** или **интервална** най-често се описва посредством **средната аритметична** или **медианата**. Разсейването при тези променливи се онагледява чрез стойността на **стандартното отклонение** или **интерквартилния размах**. За променливите измервани на **номинална** и **ординална** скала се използва единствено индикаторът относителен дял.

[4] "Централна тенденция" означава използването на стойност, която описва всички участници в извадката "като едно цяло"."Разсейване" означава колко са "разпръснати" наблюденията около централна тенденция.

Показатели за централна тенденция

Средна аритметична

Средната аритметична се изчислява в зависимост от вида на представяне на количествените данни. Те могат да бъдат изброени негрупирани числа

или таблично подредени агрегирани и групирани стойности. От своя страна групировката може да бъде извършена във **вариационен** или **интервален** ред.

Негрупирани данни

При изчисление на средната аритметична за негрупирани данни се използва формулата:

$$\bar{x} = \frac{\sum x}{n}$$

Където $\bar{x}$ е символът за средна аритметична, $\sum x$ е сборът на числовите стойности на наблюдаваната променлива, а n е броят на наблюденията.

Нека онагледим това с пример. По-долу са представени стойностите на променливата "сърдечен пулс" за 10 студента от втори курс.

[1] 90 80 89 72 73 74 78 86 84 82

Графично е представено тяхното **разпределение** на фигура фигура 2. В случая, средната аритметична представлява "средата" на това разпределение (най-издадената му част). За да я изчислим разделяме сбора от стойностите на пулса за всеки един от студентите върху техния брой. Така получаваме:

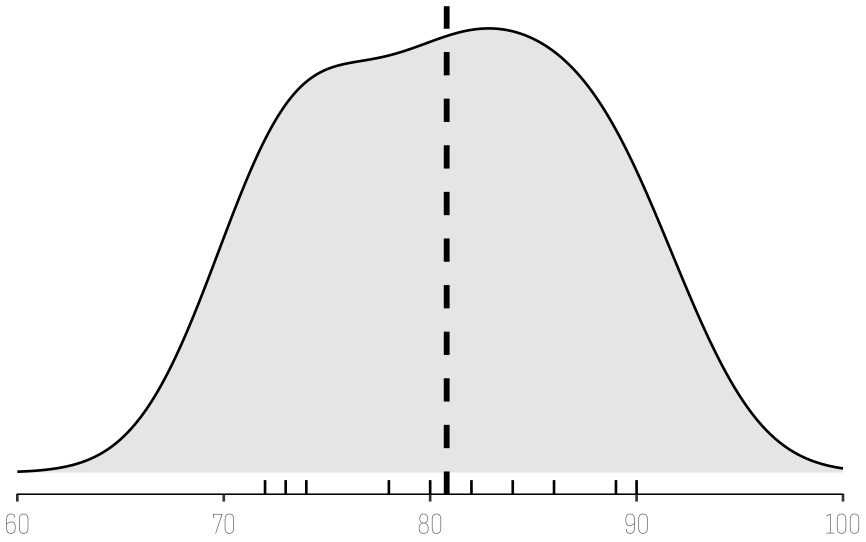
$$\bar{x} = \frac{90 + 80 + 89 + 72 + 73 + 74 + 78 + 86 + 84 + 82}{10}$$

$$\bar{x} = \frac{808}{10} = 80.8$$

Групирани данни

Интервален ред

Когато разполагаме с по-голям обем данни, удобно е тяхното представяне в групирана форма. При **интервалния ред** тази групировка се извършва чрез "разкъсване" на стойностите на изследваната променлива в равномерни и непripокриващи се стъпки - интервали. Всеки от тях се определя от долна и горна лимитираща стойност (граница). При формирането на интервалните групи се преброяват само участниците, чиято наблюдавана стойност е в границите на интервала. С други думи, при тази групировка "губим" статистическа точност, защото не знаем индивидуалните стойностите



фигура 2: Разпределение на пулса -по хоризонтала са представени наблюдавани стойности, а по вертикала, колко често са наблюдавани. С черна пунктирана линия е представена средната аритметична. Представеното разпределение се нарича нормално - има камбановидна, симетрична форма

на участниците, а само към коя група принадлежат. Пример за интервална групировка е представен в таблица 2 .

За да изчислим средната стойност за променливата пулс в посочения пример, използваме формулата:

$$\bar{x} = \frac{\sum x_u \cdot f}{\sum_f}$$

Където:

- $\bar{x}$ е символът за средна аритметична;
- x_u е средата на всеки от интервалите. Тя се изчислява като половината на сумата между горната и долната лимитираща стойност. Например, за интервала 48 - 52, средата е 50;
- f е броят на наблюденията за всеки един интервал;
- $x_u \cdot f$ е произведението между средата на интервала и броя на наблюденията в него;
- $\sum x_u \cdot f$ е сумата от всички произведения;
- $\sum_f$ е сумата от всички наблюдавани участници.

Нека представим нагледно примера в таблица 2 . Първата стъпка е определянето на средата на интервалите. Това е извършено в таблица 3 .

таблица 2: Пример за интервален ред при измерването на пулс сред 45 студенти

пулс	брой студенти
48 - 52	4
53 - 57	5
58 - 62	8
63 - 67	11
68 - 72	6
73 - 77	6
78 - 82	5

таблица 3: Стъпка 1 - Определяне на средата на всеки интервал

пулс	среда на интервала	брой студенти
48 - 52	50	4
53 - 57	55	5
58 - 62	60	8
63 - 67	65	11
68 - 72	70	6
73 - 77	75	6
78 - 82	80	5

След това изчисляваме произведенията между броя наблюдавани студенти и средата на интервала. Това е направено за всеки един интервал и представено в таблица 4 .

таблица 4: Стъпка 2 - Умножение на средата на интервала по броя на наблюдаваните студенти в него

Пулс	Среда на интервала	Брои студенти	Произведение x_u x f
48 - 52	50	4	200
53 - 57	55	5	275
58 - 62	60	8	480
63 - 67	65	11	715
68 - 72	70	6	420
73 - 77	75	6	450
78 - 82	80	5	400

Накрая сумираме всички получени произведения и ги разделяме на броя студенти

$$\bar{x} = \frac{200 + 275 + 480 + 715 + 420 + 450 + 400}{45} = \frac{2940}{45} = 65.33$$

Степенен ред

Вторият вариант за групиране на статистически данни е **степенния ред**. При него записваме всяка стойност на интересуващата ни променлива, а срещу нея нанасяме броя на наблюдаваните лица, при които сме я наблюдавали. Пример за степенен ред е представен в таблица 5 .

За изчисляването на средната аритметична, при данни в степенен ред, използваме формулата:

$$\bar{x} = \frac{\sum x \cdot f}{\sum_f}$$

Където:

- $\bar{x}$ е символът за средна аритметична;
- x е стойността на наблюдаваната променлива;
- f е броят на наблюденията за всяка една стойност;
- $x \cdot f$ е произведението между всяка от стойностите на променливата и броя на наблюденията за нея;
- $\sum x \cdot f$ е сумата от всички произведения;
- $\sum f$ е сумата от всички наблюдавани участници.

Нека представим нагледно примера в таблица 5 . Първата стъпка, за изчисление на средната аритметична, е определянето на произведенията между всяка една стойност на променливата и съответния брой на наблюденията. Това е направено в таблица 6 .

таблица 6: Пример за степенен ред при измерването на пулс сред 17 студенти

Стойност на пулса	Брой студенти	Произведение x x f
59	1	59
60	0	0
61	2	122
62	4	248
63	0	0
64	3	192
65	2	130
66	1	66
67	1	67
68	3	204

След това сумираме всички получени произведения и ги разделяме на броя студенти [5] :

таблица 5: Пример за степенен ред при измерването на пулс сред 17 студенти

Стойност на пулса	Брой студенти
59	1
60	0
61	2
62	4
63	0
64	3
65	2
66	1
67	1
68	3

[5] **Важно!** При степенния ред, за разлика от интервалния, знаем стойността на всяко едно наблюдение. Тоест не "губим" статистическа точност.

$$\bar{x} = \frac{59 + 0 + 122 + 248 + 0 + 192 + 130 + 66 + 67 + 204}{17} = \frac{1088}{17} = 64$$

Други показатели за централна тенденция

Освен средната аритметична съществуват и други показатели, които представят централната тенденция. Това са:

- **Медианата** - Медианата е тази стойност, която разделя възходящо подреден статистически ред на две равни части. С други думи 50% от наблюдаваните стойности са или по-големи или по-малки на медианата [6] [7];
- **Модата** M_0 е най-често срещаната стойност в статистическия ред и се намира чрез броене [8].

Относителен дял

Относителният дял представлява централна тенденция за променливите измервани на ординална и номинална скала. Той се изчислява чрез формулата:

$$\hat{p} = \frac{m}{n}100$$

Където:

- $\hat{p}$ е символът за относителен дял;
- m е броят на наблюденията, за които искаме да установим централната тенденция;
- n е броят на всички наблюдения в изследваната група.

Ако искаме да разберем какъв е относителния дял на 4 момичета в група от 10 студенти:

$$\hat{p} = \frac{4}{10}100 = 40\%$$

[6] Ако сме измерили кръвна на 5 пациенти със следните стойности 6.0; 6.5; **7.0**; 7.5; 8.0, стойността 7.0 е медиана - защото разделя възходящо подредения ред на две симетрични части.

[7] Ако имаме ред с четни стойности като: 6.0; 6.5; **7.0; 7.5**; 8.0; 8.5, медианата получаваме като разделим на две сбора от двете стойности в средата на реда - в случая медианата в случая е 7,25.

[8] Ако имаме следните данни за систолното кръвно налягане **120**; **120**; 130; 135; **120**; 140; 140 модата е **120** -защото представена е най-много пъти. В статистически ред е възможно да няма мода или да има две и повече моди.

Показатели за разсейване

Показателите за разсейване са мерки за разпръснатостта на стойностите спрямо централната тенденция. Те се изчисляват само за променливи измервани на **пропорционална** или **интервална** скала.

Размах

Размахът представлява разликата между **максималната и минималната** стойност в един статистически ред [9]. Ако използваме първия пример, за стойността на пулса при 10 студенти - 90; 80; 89; 72; 73; 74; 78; 86; 84; 82, размахът е

$$R = 90 - 72 = 18$$

[9] Размахът не се използва рутинно в медицинската статистика, защото е зависим само от две стойности, а понякога те са екстремални поради грешки в измерването.

Стандартно отклонение

Стандартно отклонение при негрупирани данни

Стандартното отклонение е показател за **средното ниво на разсейване** спрямо средната аритметична. Формулата за изчисление на стандартното отклонение е

$$SD = \sqrt{\frac{\sum(x - \bar{x})^2}{n}}$$

Където: - SD е символът за стандартно отклонение; - x е стойността на всяко едно наблюдение; - $\bar{x}$ е средната аритметична; - n е броят на наблюденията в изследваната група; - $x - \bar{x}$ е отклонението на всяко едно наблюдение от средната аритметична; - $\sum(x - \bar{x})^2$ е сумата от квадратите на всички отклонения.

Нека разгледаме примера с пулса на 10-тимата студенти от втори курс. Както определихме по-рано, средната стойност на пулса за тях е 80.8. Въпреки това няма нито един студент с точно тази стойност. Всеки студент се отклонява с няколко удара под или над тази средна. Това отклонение представлява разсейването (девиацията) и е първата стъпка при изчислението на стандартната грешка. Разликите между стойността на средната аритметична и стойността на пулса за всеки един студент са представени в таблица таблица 7

сърд.честота	девиация
90	9.2
80	-0.8
89	8.2
72	-8.8
73	-7.8
74	-6.8
78	-2.8
86	5.2
84	3.2
82	1.2

таблица 7: Отклонение на всеки един от изследваните спрямо средната аритметична за групата

Ако съберем девиациите на всички участници, че получим "0". Това няма да ни позволи да изчислим средното отклонение в групата - затова, във втората стъпка повдигаме всяка една от девиациите на втора степен. Изчисленията са представени в таблица 8

сърд.честота	девиация	девиация^2
90	9.2	84.64
80	-0.8	0.64
89	8.2	67.24
72	-8.8	77.44
73	-7.8	60.84
74	-6.8	46.24
78	-2.8	7.84
86	5.2	27.04
84	3.2	10.24
82	1.2	1.44

таблица 8: Девиация повдигната на квадрат

В третата стъпка - събираме стойностите на всички отклонения повдигнати на квадрат. Резултатът разделяме на броя на студентите, а за да "стандартизираме" използваното в предходния етап степенуване използваме коренуване.

$$SD = \sqrt{\frac{84.64 + 0.64 + 67.24 + 77.44 + 60.84 + 46.24 + 7.84 + 27.04 + 10.24 + 1.44}{10}}$$

$$SD = \sqrt{\frac{383.6}{10}} = \sqrt{38.36} = 6.19$$

След като вече разполагаме със стандартното отклонение, можем да изразим пулса на групата по този начин $\bar{x} = 80.8 \pm 6.19$

Стандартно отклонение при данни групирани в интервален ред

Нека разгледаме примера с пулса на 45 студента от втори курс. Изходните стойности са представени в таблица 2 . За да определим стандартното отклонение, трябва да приложим следната формула:

$$SD = \sqrt{\frac{\sum (x_u - \bar{x})^2 \cdot f}{\sum_f}}$$

Където:

- SD е символът за стандартно отклонение;
- x_u е средата на всеки от интервалите;
- $\bar{x}$ е средната аритметична;
- f е броят на наблюденията за всеки един интервал;
- $x_u - \bar{x}$ е отклонението на средата на интервала от средната аритметична;
- $\sum (x_u - \bar{x})^2 \cdot f$ е сумата от квадратите на всички отклонения, умножени по броя на наблюденията в интервала;
- $\sum_f$ е сумата от всички наблюдавани участници.

Нека разгледаме последователно как се изчислява стандартното отклонение в посочения пример.

Първата стъпка е определянето на средата на интервалите - това е показано в таблица 3 . Във втора стъпка - определяме девиацията между средата на всеки интервал и изчислената средна аритметична - **65.33**. Това е представено в таблица 9 .

таблица 9: Стъпка 2 - Определяне на девиацията

пулс	среда на интервала	Девиация $x_u - \bar{x}$	брой студенти
48 - 52	50	-15.33	4
53 - 57	55	-10.33	5
58 - 62	60	-5.33	8

пулс	среда на интервала	Девиация $x_u - \bar{x}$	брой студенти
63 - 67	65	-0.33	11
68 - 72	70	4.67	6
73 - 77	75	9.67	6
78 - 82	80	14.67	5

В трета стъпка - повдигаме девиациите на квадрат, както е показано в таблица 10 .

таблица 10: Стъпка 3 - Повдигане на девиациите на втора степен

пулс	Среда на интервала	Девиация $x_u - \bar{x}$	$(x_u - \bar{x})^2$	брой студенти
48 - 52	50	-15.33	235	4
53 - 57	55	-10.33	106.7	5
58 - 62	60	-5.33	28.40	8
63 - 67	65	-0.33	0.11	11
68 - 72	70	4.67	21.8	6
73 - 77	75	9.67	93.7	6
78 - 82	80	14.67	215,2	5

След това умножаваме всяка една от получените стойности с броя на наблюденията в съответния интервал. Това е представено в таблица 11 .

таблица 11: Стъпка 4 - Умножаване на квадрата на девиациите по броя студенти

пулс	среда на интервала	$(x_u - \bar{x})^2$	брой студенти (f)	$(x_u - \bar{x})^2 \cdot f$
48 - 52	50	235	4	940
53 - 57	55	106.7	5	533.5
58 - 62	60	28.40	8	227.2
63 - 67	65	0.11	11	1.21
68 - 72	70	21.8	6	130.8
73 - 77	75	93.7	6	562.2
78 - 82	80	215,2	5	1076

Накрая сумираме всички получени произведения, разделяме ги на броя студенти и коренуваме

$$SD = \sqrt{\frac{940 + 533.5 + 227.2 + 1.21 + 130.8 + 562.2 + 1076}{45}}$$

$$SD = \sqrt{\frac{3470.91}{45}} = \sqrt{77.13} = 8.78$$

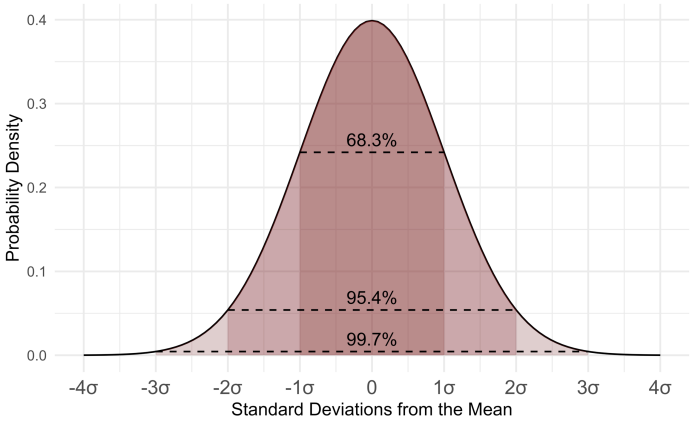
> **!** Важно
>
> В основата на стандартното отклонение можем да изградим **предикативен интервал**. Това е възможно, само ако изследваната променлива е **нормално** разпределена. За да "предскажем" какъв брой от изследваните имат стойност в конкретен интервал използваме правилото на трите сигми:
>
> - 95 % от наблюденията **в извадката** имат стойност на променливата в интервала между средната аритметична + или - **две** стандартни отклонения;
> - 68% от наблюденията **в извадката** имат стойност на променливата в интервала между средната аритметична + или - **едно** стандартно отклонение;
> - 99.7% от наблюденията **в извадката** имат стойност на променливата в интервала между средната аритметична + или - **три** стандартни отклонения.
>
> *Нормално разпределена величина - означава, че имаме много наблюдения със стойности близки до средната аритметичната, която съвпада с медианата и модата. При построяването на графика - разпределението е камбановидно и симетрично.*
>
> *Ако сме изчислили теглото на извадка от 1000 души и сме установили че $\bar{x} = 80 \pm 10$ и приемем променливата за **нормално разпределена** можем да твърдим, че 95 % от участниците (950 изследвани) имат тегло в диапазона $80 \pm 2 \cdot SD = 80 \pm 20$ или от 60 до 120 кг. Останалите 2,5 % от останалите 50 участници (25) ще имат тегло над 110 кг, а 25 души (2,5 %) ще имат тегло под 60 кг.*

Илюстративно нормалното разпределение е представено на фигура фигура 3 .

Интерквартилен размах

Понякога средната аритметична не е подходяща за измерване на централната тенденция. Такъв случай е разгледан по-долу, когато използваме данните за продължителността на хоспитализацията на 20 пациенти.

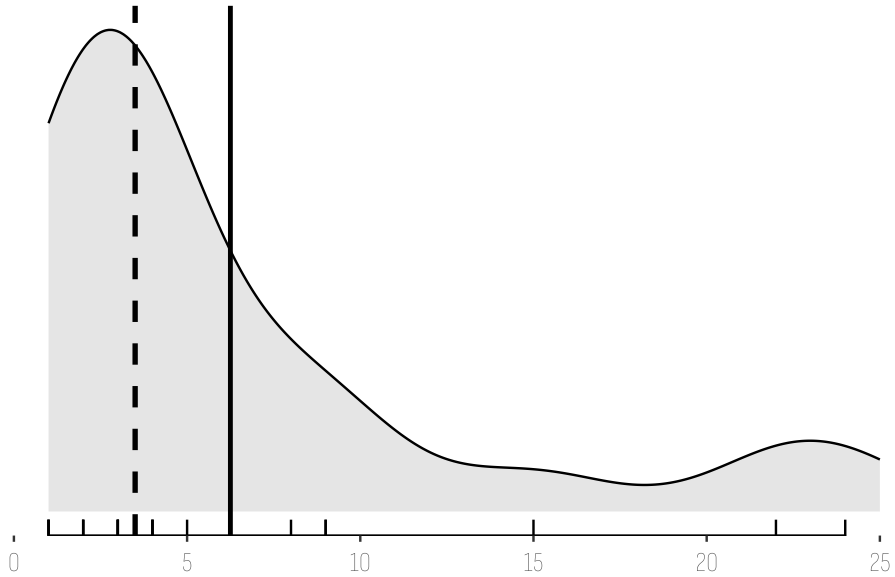
[1] 1 2 1 2 3 1 5 4 3 4 2 3 8 9 3 4 9 15 22 24



фигура 3: Нормално разпределение

По-голямата част от включените в статистическия ред пациенти са прекарали между 1 и 6 дни в болница. Въпреки това, ако изчислим средната аритметична за цялата група получаваме 6.25 дни. Това число не представя добре централната тенденция, тъй като само 5 от 20-те пациенти са пролежали в болница за повече от 6 дни. Техните стойности се наричат екстремални, а средната аритметична силно повлияна от тях.

Ако нанесем броя на дните по хоризонтала, а броя на пациентите по вертикала ще получим разпределението на променливата "продължителност на хоспитализацията" - тя е представена на фигура 4.



фигура 4: В този случай, средната аритметична (непрекъснатата линия) е изместена към екстремалните стойности. По-добър показател е медианата - тя е 3,5 дни (прекъснатата линия) и по-добре представя централната тенденция в данните.

Представеното в фигура 4 разпределение **дясно изтеглено**. С други думи, по-голямата част от наблюденията се намират в лявата част на средната аритметична, а екстремалните 5-ма пациенти издърпват "опашката" на разпределението на дясно. В случаи като тези, медианата е по-добър показател за централна тенденция.

За да разберем какво е разсейването спрямо медианата използваме показателя **интерквартилен размах**. Подобно на стандартното отклонение, той онагледява разсейването на наблюденията спрямо медианата. Показателят се изчислява по формула:

$$IQR = Q3 - Q1$$

Където:

- IQR е символът за интерквартилен размах;
- $Q3$ е стойността на 3-тия квартил [10];
- $Q1$ е стойността на 1-вия квартил.

[10] "Квартилите" са стойностите на наблюденията, които разделят статистическия ред на 4 равни части.

За да разберем какво представлява интерквартилния размах, нека първо представим данните за болничния престой във възходящ ред:

[1] 1 1 1 2 2 2 2 3 3 3 3 4 4 4 5 8 9 9 15 24

Ако разделим тези 20 пациенти в 4 групи (по 5-ма) ще установим стойностите на 4-те квартила. В случая 5-тия пациент, пролежал 2 дни, е на границата между първата и останалите три четвъртини на статистическия ред. Стойността, на неговото пролеżаване представлява 1-вия квартил $Q1$. Стойността на пролеżаването на 15-тия участник (4 дни) ограничава трите първи четвъртини от последната група пациенти - това е стойността на 3-тия квартил $Q3$.

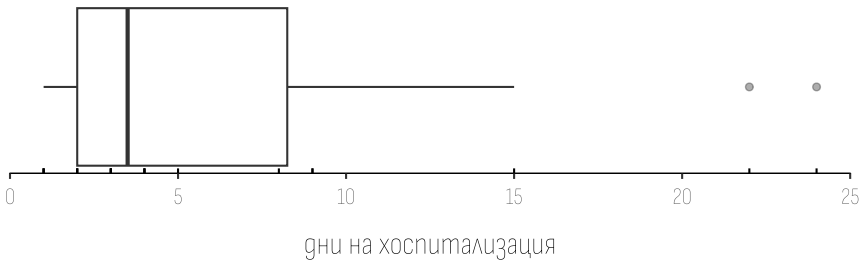
След като сме определили квартилите прилагаме формулата. Интерквартилният размах е

$$IQR = Q3 - Q1 = 4 - 2 = 2$$

За графично представяне на интерквартилния размах използваме боксплот диаграмата представена на фигура 5 .

С помощта на интерквартилния размах можем да определим кои стойности са "екстремални" [11]

[11] Екстремални са всички стойности по-високи от $Q3 + 1.5 \cdot IQR$ или по-ниски от $Q1 - 1.5 \cdot IQR$. В посочения пример такива са стойностите над $4 + 1.5 \cdot 2$ и под $2 - 1.5 \cdot 2$.



фигура 5: Боксплотът се описва като кутия с мустаци. Лявата страна на кутията представя първия, а дясната страна третия квартил. В средата на кутията с черна линия хоризонтална линия е представена медианата.

Инферентна статистика

Показателите централна тенденция и разсейване в извадката се наричат **статистики**. При инферентния анализ, тези индикатори се използват за да се направи заключение за интересуващите ни **параметрите** в генералната съвкупност [12].

> **!** Статистика и параметър
>
> Статистиката изчислена в извадка се използва за да се установят **вероятните граници** на параметъра в генералната съвкупност

За да се определи интервала, в който с определена сигурност се намира стойността на интересуващият ни параметър, трябва първо да изчислим стандартната грешка.

[12] Средният пулс на групата от 10 студенти представлява **статистика**, а средният пулс на генералната съвкупност - студентите от II курс представлява **параметър**.

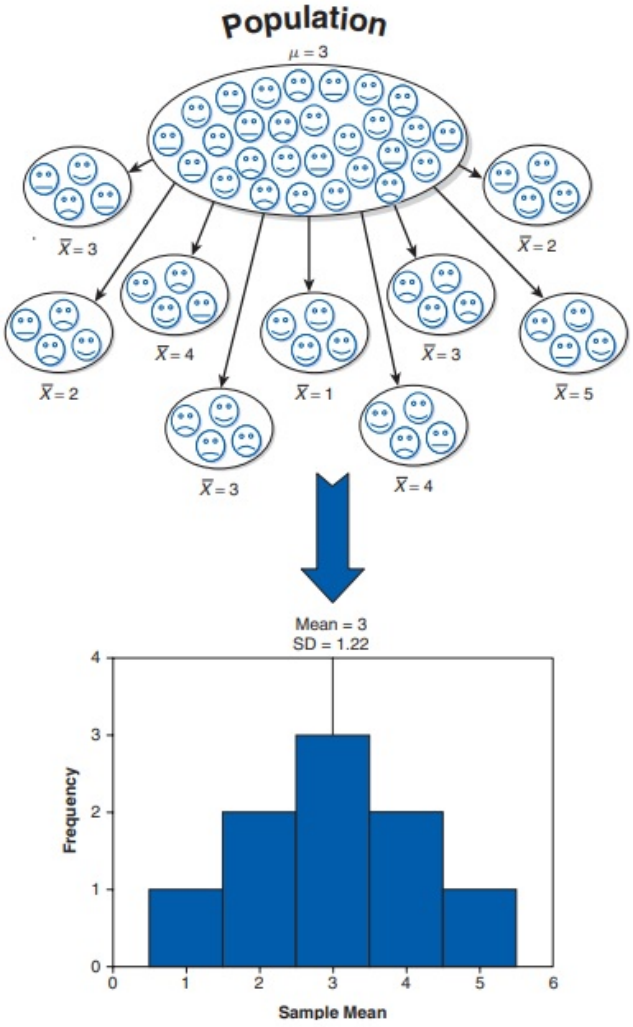
Стандартна грешка на средната аритметична (SEM)

Нека си представим, че имаме генерална съвкупност от 200 студенти със стойност на средната аритметична равна на 3. Ако направим 9 случайни извадки от нея (всяка по 10 студенти), можем да определим средната аритметична във всяка една от тях. Ако нанесем 9-те статистики на една графика ще получим нормално разпределение от извадковите средни аритметични. Това е представено в фигура 6.

Колкото повече извадки правим, толкова повече средната от всички извадкови статистики ще е по-близо до истинската средна за генералната съвкупност. Това е **закона на големите числа**. **Стандартната грешка** ни позволява да оценим точността на средната аритметична за всяка извадка, спрямо неизвестен параметър на генералната съвкупност.

За да изчислим стандартната грешка прилагаме формулата:

$$SE_m = \frac{SD}{\sqrt{n}}$$



фигура 6: Разпределение на средни аритметични на 9 извадки от генералната съвкупност

Където:

- SE_m е символът за стандартна грешка на стандартна грешка;
- SD е символът за стандартно отклонение;
- n е броят на наблюденията в извадката.

Грешката зависи от стандартното отклонение - колкото по-голямо е то, толкова по-неточна е оценката. Също така от значение е обемът на извадката n - колкото повече наблюдение сме включили в извадката толкова по-малка е грешката.

Ако използваме примера с 10-мата студенти със среден пулс от 80.8 удара в минута и стандартно отклонение от 6.16 уд./мин, можем да изчислим стандартната грешка като заместим:

$$SE_m = \frac{SD}{\sqrt{n}} = \frac{6.19}{\sqrt{10}} = \frac{6.19}{3.16} = 1.95$$

Във втория пример, с данните за 45-мата студенти в интервален ред със средната аритметична 65.33 уд./мин и стандартното отклонение 8.78 уд./мин. Стандартната грешка е

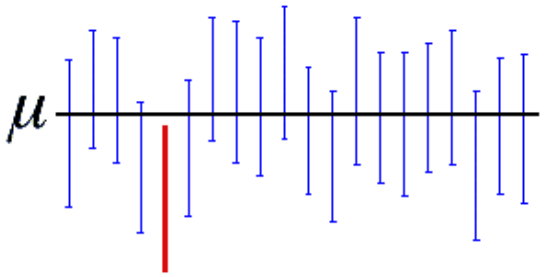
$$SE_m = \frac{SD}{\sqrt{n}} = \frac{8.78}{\sqrt{45}} = \frac{8.78}{6.71} = 1.31$$

Интервал на доверителност средна аритметична

Интервалът на доверителност представлява оценка на параметъра на генералната съвкупност. Той ограничава възможните стойности на известният параметър определена вероятност (сигурност). Тази степен на сигурност се определя от изследователя - в медицината най-често се използват 95% , 99% или 90% вероятност.

За да се онагледи концепцията за интервал на доверителност се използва фигура 7 .

Хоризонталната черна линия представлява "неизвестния" параметър в генералната съвкупност - в нашия пример това е средната пулсова честотата на II курс. Вертикални линии представляват 95%-товите интервали на доверителност изградени в основа на 20 извадки. При 19 от тях това интервалът обхваща стойността на истината средна за генералната съвкупност (сини линии). При един интервал (5%) интервалът е неточен (в червено) и не съдържа истинската стойност на средната сърдечна честота в курса. Не е възможно изграждането на 100 % интервал ! Статистиката



фигура 7: 95% интервал на доверителност сред 20 извадки

единствено ни дава възможността да намалим несигурността си по отношение на неизвестния параметър.

За да конструираме интервала на доверителност използваме формулата:

$$CI = \bar{x} \pm z \cdot SE_m$$

Където

- CI е символът за интервал на доверителност;
- $\bar{x}$ е средната аритметична;
- z е гаранционният множител;
- SE_m е стандартната грешка.

Гаранционен множител z е константа, която зависи от това избраното ниво на сигурност с искаме да определим интервалът на доверителност. Най-често избираното ниво в медицината е **95%**, за което стойността на z е 1,96 [13].

Нека използваме примера за 10 -те студенти със среден пулс 80,8 и стандартна грешка от 1.95. 95% -овия интервал на доверителност се изчислява като:

$$95\% CI = 80.8 \pm 1.95 \cdot 1.96 = 80.8 \pm 3.82$$

В основа на нашата извадка, можем да кажем, че с **95% сигурност, средната сърдечна честота на курса се намира в диапазона между 76.98 и 84.62 уд./мин**.

[13] Ако искаме да конструираме 99% интервал, стойността z е 2.58, при 90% интервал - 1.65

Стандартна грешка и интервал на доверителност за пропорция

Стандартната грешка може да бъде изчислена за качествени променливи измервани на номинална и ординални скали. Нека вземем например за параметър относителния дял на момичетата в 2-ри курс. Да приемем, че нямаме достъп до генералната съвкупност и трябва да направим заключение в основата на една студентска група от 12 студента, 6 от които са момичета.

Първо - изчисляваме относителния дял (статистика) в извадката

$$\hat{p} = \frac{6}{12} = 50\%$$

След това изчисляваме стандартната грешка

$$SE_p = \sqrt{\frac{\hat{p} \cdot (100 - \hat{p})}{n}}$$

$$SE_p = \sqrt{\frac{50 \cdot 50}{12}} = \sqrt{\frac{2500}{12}}$$

$$SE_p = \sqrt{208} = 14.4\%$$

Накрая определяме 95%-овия интервал на доверителност за параметъра в генералната съвкупност:

$$CI = \bar{x} \pm z \cdot SE_m$$

$$95\% CI = 50\% \pm 1.96 \cdot 14.4\%$$

$$95\% CI = 50\% \pm 28.2\%$$

Изводът е, че с **95%-ова сигурност** можем да твърдим, че относителният дял на момичетата в 2-ри курс е в интервала между 21.8% и 78.2%.

Важно

Ако изберем да използваме ниво на сигурност 99% умножаваме по по-висока стойност на гаранционния множител - това ще доведе до разширение на ширината на интервала на доверителност. Обратно, ако изберем да използваме ниво на сигурност 90% умножаваме по по-ниска стойност на гаранционния множител - това ще доведе до намаление на ширината на интервала на доверителност.

02-Упражнение 2023/2024

Домашна работа

ас.д-р Костадин Костадинов

Задачи

Задача 1.

След провеждане на тренировъчен режим с група спортисти в края на периода е измерено теглото в килограми. Получени са следните стойности: 64, 76, 68, 84, 80. 72, 74, 72 кг. Да се определи средното тегло на групата и границите на 95% от интервал на доверителност за средното тегло на спортистите в популацията.

Задача 2.

Изследвана е съобразителността на 40 студента при работа с компютри по определена психологична задача. Да се изчисли средното време, употребено от студентите за решаването и при следните резултати. Да се определят границите на 95% - овия доверителен интервал за средната аритметична. Времето е количествена променлива.

Употребено време (мин)	Брой студенти
3	1
4	3
5	4
6	7
7	8
8	9
9	8
общо	40

Задача 4

Дадена е променливата "диагноза" с нейните разновидности. Преобразувайте категорийната променлива в дихотомна и изчислете границите на 95%-овия интервал на доверителност за относителния дял на пациенти с хепатит от популацията.

Диагноза	Брой пациенти
Колит	10
Гастрит	18
Апендицит	22
Язва	29
Хепатит	38
Общо	117

Задача 5

По случаен принцип са подбрани две извадки с обем съответно n = 320 и n = 540 индивиди с белодробни проблеми. Изследваните случаи от първата извадка са третирани с антибактериален препарат "А" без витамини, докато в терапията на втората извадка са включени и витамини. Случаите с подобрение в първата група са 98, а във втората група 231. Да се изчислят границите на 95% в доверителен интервал за относителния дял на случаите с подобрение в двете извадки.

Терапия	Всички пациенти	Случаи с подобрение
А без витамини	320	98
А с витамини	540	231

Тестови

1. Кой от изброените вариационни редове има мода?

a) 1, 4, 5, 7, 9, 11, 12, 13
b) 2, 4, 5, 7, 9, 11, 12, 13
c) 1, 4, 5, 7, 9, 11, 11, 13
d) 1, 4, 5, 7, 9, 11, 12, 14

2. Кое от следните твърдения е грешно?

a) Степенните вариационни редове водят до по-прецизни крайни оценки в сравнение с интервалните.
b) Количествените променливи винаги могат да бъдат преобразувани в качествени.

c) Променливата пол се измерва на номинална скала.
d) Променливата ръст при раждане се измерва на интервална скала.

3. В случай на дясно изтеглено разпределение:

a) Средната аритметична е по-голяма по стойност от модата и показателят за асиметрия е положителен.
b) Средната аритметична е по-малка по стойност от модата и показателят за асиметрия е положителен.
c) Средната аритметична е по-голяма по стойност от модата и показателят за асиметрия е отрицателен.
d) Средната аритметична е по-малка по стойност от модата и показателят за асиметрия е отрицателен.

4. Доверителна вероятност от 90% означава:

a) 90 от 100 доверителни интервала биха довели до грешно заключение за популационния параметър.
b) 100 от 1 000 доверителни интервала биха довели до грешно заключение за популационния параметър.
c) Използваме гаранционен множител Z=1.96 при изчисляване на границите на доверителния интервал.
d) Вероятността за вярно заключение относно изучавания популационен параметър е 10%.

5. Кой от изброените отговори съдържа само променливи, които могат да бъдат измерени на интервална или пропорционална скала?

a) Кръвна група, систолно кръвно налягане, стадий на заболяване
b) Ниво на кръвна захар, тегло, индекс на телесна маса
c) Индекс на телесна маса, кръвна група, семеен статус
d) Систолно кръвно налягане, индекс на телесна маса, пол

03-Упражнение 2023/2024

Тест на хипотеза. T-тест.

ас.д-р Костадин Костадинов

Важни правила от дескриптивната и инферентна статистика

1. За да се опише статистически една извадка е необходимо да:
    - се определи **централна тенденция** на всички наблюдавани променливи при разглеждането им като едно цяло [1];
    - се определели нивото на **разсейване** спрямо тази тенденция [2];
2. Променливите биват **количествени** и **качествени**:
    - *Количествените* се измерват на силни скали [3];
    - *Качествени* се измерват на слаби скали [4]:
        - Ако признакът е подредим - ординална скала (нп. *стадии на онкологично заболяване*);
        - Ако признакът приема само две възможни стойности - дихотомна скала (нп. *пол - мъж/жена*);
3. Когато количествени променливи са **нормално разпределение** [5], те се описват със средната аритметична $\bar{x}$ и стандартното отклонение SD;
4. При асиметрично разпределение (ляво или дясно изтеглено) по-подходящи са медианата Me и интерквартилния размах IQR;
5. Стандартното отклонение:
    - Е измерител на средното ниво на вариабилност около средната аритметична;
    - Може да бъде "0" при наблюдения с едни и същи стойности;
    - Намалява с увеличение на наблюденията в извадката;
    - Е високо, ако наблюдаваните стойности са много разпръснати около средната аритметична;
6. Интерквартилния размах:

[1] Показатели за централна тенденция са *средната аритметична*, *медианата* или *относителен дял*

[2] Показатели за разсейване са *стандартното отклонение* и *интерквартилния размах*

[3] Като *пропорционална* или *интервална*;

[4] Като *номинална* и *ординална*;

[5] Нормалното разпределение е симетрично, камбановидно.

- Представлява нивото на вариабилност спрямо медианата.
- Се представя чрез диаграмата **boxplot**.
  - Тя представлява кутия с "мустаци" - двете и страни съответстват на стойностите на първия Q1 и третия Q3 квартил, а средната линия, разделяща кутията, е медианата.
  - Интерквартилният размах е разликата между стойностите на третия и първия квартил;
  - "Мустаците" от двете страни на кутията определят вида на разпределението [6];

7. В основа на резултатите от **извадката** (статистики) се изгражда интервал от стойности, в който с определена степен на сигурност, се намира параметърът за генералната съвкупност [7].
   - Първата стъпка при построяването на доверителния интервал е определянето на **стандартната грешка**
     - Тя зависи от броя наблюдения - увеличаването на броя води до намалява стандартната грешка;
     - Тя зависи от стандартното отклонение - увеличаването на стандартното отклонение води до увеличаване на стандартната грешка;
   - Широчината на **доверителния интервал** зависи от *стандартната грешка* и от избраното *ниво на сигурност* [8];
   - При равни други условия, ако увеличим степента на сигурност от 90 до 99% широчината на интервала нараства;
   - При високи стойности на стандартна грешка интервалът на доверителност също е широк;

[6] Ако мустаците са равни - разпределението е симетрично, ако десния мустак е по-дълъг от левия - разпределението е дясно изтеглено, ако левият мустак е по-дълъг от десния - разпределението е ляво изтеглено.

[7] Пример за "статистика" е средната аритметична за ръста в *една извадка*, параметър е средната аритметична за ръста в *генералната съвкупност*.

[8] В медицината най-често се използва $95\% CI$

Тест на хипотези

Освен да опишем една извадка и да направим предположение за това каква е стойността на показателя в генералната съвкупност, статистиката се използва и за тестване на хипотези. Хипотезата е научно предположение за наличие на връзка между две или повече изследвани променливи. Хипотезите биват два вида - нулева и алтернативна.

При тестването на хипотези се определя колко е вероятно една взаимовръзка да бъде наблюдавана в резултат на случайност. Тази вероятност се бележи със символа p. В класическата теория на статистиката се *"тества"* единствено **нулевата хипотеза** - приемаме, че тя е вярна, ако вероятността p е равна или по-голяма от 5% (0.05). От друга страна, ако p е под 5% (0.05) отхвърляме нулевата хипотеза - и приемаме за вярна алтернативната [9].

[9] Някой изследователи не приемат зададеното ниво за p от 0.05 (5%). Алтернативно други предложени гранични стойности за отхвърляне на нулевата хипотеза са p < 0,01 (1%) или p < 0.1 (10 %).

Определения

Нулевата хипотеза H_0 е винаги **отричащо** твърдение за наличието на връзка между две или повече явления.[a]

Алтернативната хипотеза H_1 е твърдение **противоположно** на нулевата хипотеза. [b].

[a] H_0 е твърдението *"Новият медикамент за хипертония **не** намалява стойностите на кръвното налягане, ако има подобно намаляване, то е наблюдавано **случайно**"*

[b] H_1 е твърдението *"Новият медикамент за хипертония намалява стойностите на кръвното налагане, това намаление е **не в резултат на случайност**".*

Пример

Тестваме ново лекарство за хипертония, като избираме две групи пациенти (извадки). Случайно избираме едната от тях, която лекуваме с новия медикамент, а на другата прилагаме плацебо таблетка, която не съдържа лечебно вещество. В края на изследването сравняваме средните стойности на кръвното налягане в двете групи пациенти. Нека приемем, че в експерименталната група кръвното налягане е средно 10 mmHg по-ниско от това в плацебо групата.

H_0 Нулевата хипотеза гласи *"Не съществува връзка между приема на медикамента и стойността на кръвното налягане. Тази разлика от 10 mmHg е в следствие на случайността."*
H_1 Алтернативната хипотеза е противоположна *"Има връзка между медикамента и понижаването на кръвното налягане. Тази разлика от 10 mmHg **НЕ** е в резултат на случайността.*

Чрез статистически методи определяме **колко е вероятно** да получим подобна разлика случайно (p). Ако приемем, че в нашия пример p е 0.09 (9%), това означава, че ако новият медикамент действително не работи и повторим експеримента с други пациенти 100 пъти, такава или по-голяма разлика в кръвното налягане ще се наблюдава случайно в 9 от тези 100 опита. В случая това е "прекалено много" за да отхвърлим нулевата хипотеза. Обратно, ако сме изчислили, че p е по-малко от 0.05 (9%), това означава, че ако в действителност новият медикамент не работи и повторим експеримента с други пациенти 100 пъти, такава или по-голяма разлика в кръвното налягане ще се наблюдава случайно в по-малко от 5 в тези 100 опита. В случая това е достатъчно за да отхвърлим нулевата хипотеза и приемем алтернативната.

Видове грешки

Определянето на "вероятността за случайност" p зависи от множество фактори. Понякога е възможно да направим "погрешен" статистически извод, следствие на това, че сме наблюдавали много малък брой участници или сме ги подбрали по неправилен начин. При изграждане на извод в основа на тестването на хипотези са възможни два вида грешки [10] представени в таблица 1

таблица 1: Видове грешки при тестването на хипотеза

	H_0 е **вярна**	H_0 **е грешна**
Отхвърляме H_0	Грешка от първи род α	Правилно решение
Приемаме H_0	Правилно решение	Грешка от втори род β

- Грешката от първи род α е от особено значение на статистическия анализ. При нея се приемат за "научни открития" явления, които в действителност **не** са верни. Граничната стойност на вероятността p, под която отхвърляме нулевата хипотеза, представлява допустимото ниво на α грешката.
- Грешка от втори род β е свързана основно с използвания статистически тест и броя на включените в изследването наблюдения. В основа на тази грешка не сме способни да докажем действително съществуващи връзки.

[10] Двата вида грешки са взаимосвързани (**но не пропорционално**) - ако намалим вероятността за грешка от първи род, като променим гранична стойността на p от 0.5 на 0.1, ще се увеличим възможността за грешка от втори род. Обратно, ако приемем, че отхвърляме нулевата хипотеза при ниво от 0.1, вместо 0.5 - увеличаваме възможността за грешка от първи род α, но намаляваме възможността за грешка от втори β.

Мощност

Мощността е характеристика на статистическия тест и представлява възможността докажем връзка между две явления, когато действително такава съществува. Мощността е равна на $1 - \beta$ и зависи от:

1. Риска за грешка от 1-ви род α. Увеличаването на допустимото ниво на риск за грешка от 1-ви род α от 0.5 до 0.1, намалява грешката от втори род β и увеличава мощността.
2. Размер на "терапевтичния ефект". При много силни взаимовръзки в статистическите явления, тестът има по-голяма мощност [11].
3. Хомогенността на групата. Колкото по-хомогенно е представена статистическата връзка в изследваните индивиди, толкова по-голяма е мощността на теста.
4. Обемът на извадките. Високия брой наблюдения в извадката води до по-малко стандартно отклонение и стандартна грешка. От своя страна по-малката стандартна грешка повишава мощността на статистическия тест.

[11] В ситуация, в която "новият" медикамент действително е много силен и драстично намаля кръвното налягане, разликата между с плацебо групата ще е голяма, а вероятността да сме я наблюдавали случайно - ниска.

Видове статистически тестове

Статистическите тестове са математическите "инструменти", с които се определя числено вероятността за случайно установяване на взаимовръзка между две или повече променливи. Изборът на тест зависи от:

1. Вида на променливата, която изследваме [12]
2. Броя на изследваните групи [13]
3. Дизайна на изследването [14]

В таблица 2 и таблица 3 са представени основните статистически тестове спрямо разгледаните по-горе признаци.

таблица 2: Видове статистически тестове при независим дизайн (еднократно изследване на групите)

Групи	Параметричен тест (**количествени**)	Непараметричен тест (**количествени**)	Непараметричен тест (**качествени**)
1	Едногрупов t тест	Wilcoxon тест	χ^2 за съответствие
2	t тест	Mann Whitney U тест	χ^2 тест
≥ 3	ANOVA	Kruskal Wallis H тест	χ^2 тест

таблица 3: Видове статистически тестове при зависим дизайн (едни и същи групи са изследвани многократно във времето)

Групи	Параметричен тест (**количествени**)	Непараметричен тест (**количествени**)	Непараметричен тест (**качествени**)
2	RM - t test	Wilcoxon тест	McNemar тест
≥ 3	RM-ANOVA	Friedman тест	TMcNemar тест

Статистическа и клинична значимост

Потвърждаването на статистически значима връзка между две променливи не означава, че тя е и клинично значима. Ако приемем, че нов медикамент понижава стойностите на кръвното налягане статистически значимо, но с минимален ефект от 1 mmHg не можем да твърдим, че той приложим за болшинството пациенти страдащи от това заболяване.

[12] Важно е да определим типа на променливите за които предполагаме връзка - възможни са всички комбинации между качествени и/или количествени променливи. Ако изследване връзка с участие на количествени променливи, които са **нормално разпределени** използваме **параметрични** тестове, за всички останали - **непараметрични**.

[13] Възможно е да тестваме резултатите получени от една единствена група спрямо предполагаема от литературата стойност. Повечето изследвания сравняват две групи - експериментална и контролна, а други включват допълнителни множество изследвани групи.

[14] Трябва да се определи дали изследваните групи са "независими" - тоест такива, които се изследват еднократно или "зависими" - при които провеждаме изследване на едни и същи единици в два или повече момента във времето.

⚠ Определения

Статистическата значимост представлява математическа оценка на вероятността да получим подобен или по-висок резултат в следствие на случайност.

Клинична значимост представлява оценка на ефектът на статистическата взаимовръзка в реалната клинична практика [a].

[a] Нов медикамент, може да е "статистически значим" като редуцира кръвното с 2 mmHg, но това не е клинично значим ефект нито пациента, нито вие като лекар.

T-тест

T тестът установява наличието на **разлика** в една нормално разпределена количествена величина в две независими една от други групи.

Условия за използване на T тест [15]

1. Количествената величина в двете изследвани извадки е нормално разпределена и има приблизително еднаква вариабилност (стандартно отклонение);
2. Включени са поне 30 единици във всяка от двете групи;
3. Извадките са представителни по отношение на генералната съвкупност.

[15] T тестът се използва основно за количествени признаци, но е разработен и модифициран вариант за качествени променливи измерени чрез относителен дял.

T-тест за количествени признаци

T- тестът се прилага при следването на следните 7 стъпки:

1. Дефиниране на нулева хипотеза H_0 и алтернативна хипотеза H_1;
2. Определяне на ниво на значимост α;
3. Определянето на средната аритметична и стандартното отклонение в двете групи;
4. Изчисляване на стандартната грешка;
5. Изчисляване на t статистиката;
6. Определяне на p стойността;
7. Статистическо заключение.

За да представим стъпките нека разгледаме следния *пример.* Проведено е проучване с цел да установим дали нов медикамент - бета блокер е ефективен за намаляване на сърдечната честотата [16]. Изследователката компанията е

[16] При пациенти със сърдечно-съдови заболявания, пулс около 60 е оптимален, защото позволява на сърцето да си "почива" по-дълго във фазата на диастола, по време на която се храни с кръв от коронарните артерии.

извършила първоначален експеримент с 80 случайно избрани пациенти - 40 са получили новия медикамент, а останалите 40 плацебо таблетка. След 4 месеца са измерени стойностите на сърдечната честота на всеки един участник. Данните са представени в таблица 4 и в таблица 5

Експериментална група
58, 58, 58, 59, 60, 60, 60, 60, 61, 61, 62, 62, 63, 64, 64, 64, 64, 65, 65, 65, 65, 66, 66, 66, 66, 67, 67, 67, 68, 68, 68, 69, 69, 70, 70, 70, 71, 71, 71, 72

таблица 4: Стойности на сърдечна честота на пациентите в експерименталната група

Контролна група
68, 68, 70, 71, 71, 72, 72, 74, 74, 75, 75, 76, 76, 78, 78, 79, 80, 80, 81, 82, 82, 82, 83, 84, 84, 85, 85, 86, 88, 89, 89, 90, 91, 92, 93, 93, 93, 94, 94, 96

таблица 5: Стойности на сърдечна честота на пациентите в контролната група група

Дефиниране на нулева хипотеза H_0 и алтернативна хипотеза H_1

Нулева хипотеза Няма разлика между средната сърдечна честота в двете групи пациенти. В случай, че такава се наблюдава, това е в резултат на случайност [17].

Алтернативна хипотеза Има значителна разлика в средната сърдечна честота между групата ползваща лекарство и тази приемаща плацебо [18]

[17] Това е тестваната хипотеза. Тя предполага, че независимо в коя група е попаднал пациента съществена разлика в пулса не се наблюдава. Дори и да има такава, това ще е в резултат на случайността (а не на медикамента).

[18] Алтернативната хипотеза е противоположна на нулевата нулевата. Според нея разлика в стойностите на пулса има и причината за това е новия медикамент.

Определяне на риска за грешка α

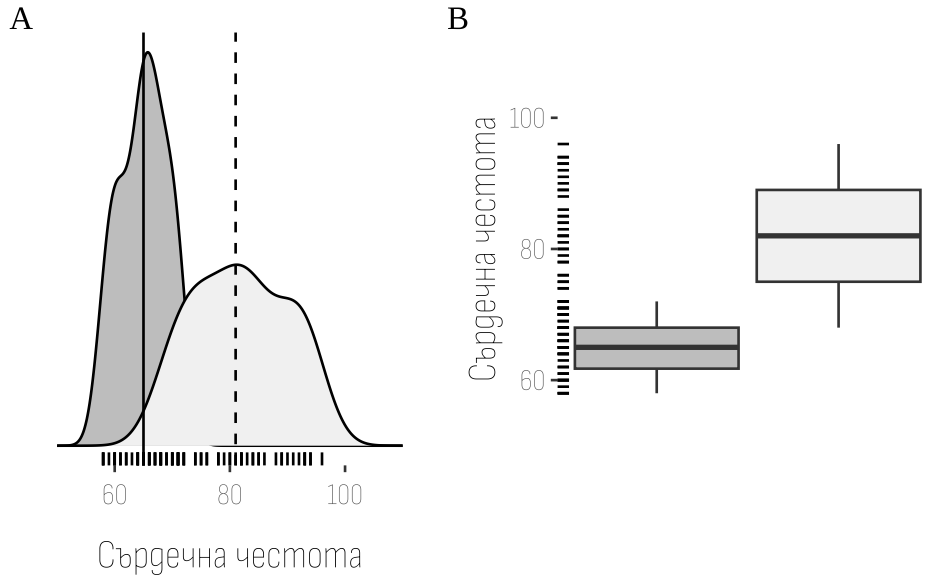
Както посочихме по-рано α в медицинските изследвания е **0.05 (5%)**. С други думи можем да "отхвърлим" нулевата хипотеза, само ако вероятността да наблюдаваме явлението в резултат на случайност е под риска за грешка от 5 %.

Определянето на средната аритметична и стандартното отклонение в двете групи [19]

От фигура 1, както и от таблица 6 можем да добием представа за това как е разпределена величина пулс в двете изследвани групи. Пациентите лекувани с бета блокер се характеризират с среден пулс от 65 уд/мин и стандартно отклонение от 4,08 уд/мин, докато при пациентите приемащи плацебо

[19] Формулите и методът за определяне на средната аритметична и стандартното отклонение са представени в упражнение 2.

средният пулс е близо 82 уд/мин, а стандартното отклонение 8.16 уд/мин. Наблюдаваната разликата между средните величини в двете групи е почти 17 уд/мин.



фигура 1: В групата с бета блокер (тъмно сиво) средният пулс е 65уд/мин (непрекъсната линия), докато в контролната група (светлосиво) пулсът е 81 уд/мин (прекъсната линия). Добиваме и представа и за хомогеността. Пулсът на пациентите с бета блокер варира малко, а стойностите са близко около средната аритметична, докато в плацебо групата вариацията е по-голяма - разпределението по-широко.

характертистика	Бета блокер	Плацебо
Средна аритметична	65.00	81.83
SD	4.08	8.16
Медиана	65.00	82.00
IQR	6.25	14.00

таблица 6: Описателна статистики на променливата сърдечен пулс в двете групи

Изчисляване на t-статистиката

За да достигнем до извод колко е вероятно тази разликата в пулса да е случайна, трябва да изчислим стойността t като използваме формула

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{{SE_{m_1}}^2 + {SE_{m_2}}^2}}$$

Където:

- $\bar{x}_1$ е средната аритметична на пулса в групата с бета блокер;

- $\bar{x}_2$ е средната аритметична на пулса в контролната група;
- SE_{m_1} е стандартната грешка на средната аритметична в групата с бета блокер;
- SE_{m_2} е стандартната грешка на средната аритметична в контролната група;
- $|\bar{x}_1 - \bar{x}_2|$ е абсолютната стойност на разликата между средните аритметични;
- $\sqrt{{SE_{m_1}}^2 + {SE_{m_2}}^2}$ е корен квадратен от сумата на квадратите на стандартните грешки на средните аритметични;
- t е стойността на t-статистиката.

Изчисляване на стандартните грешки

За да изчислим стандартната грешка за всяка от групите прилагаме вече позната формула от упражнение 2

За групата с бета блокер:

$$SE_{m_1} = \frac{SD_1}{\sqrt{n_1}}$$

$$SE_{m_1} = \frac{4,08}{\sqrt{40}} = \frac{4,08}{6,32} = 0,75$$

За групата с плацебо [20]

$$SE_{m_2} = \frac{SD_2}{\sqrt{n_2}}$$

$$SE_{m_2} = \frac{8,16}{\sqrt{40}} = \frac{8,16}{6,32} = 1,29$$

[20] Тук е момента да припомним, че плацебо групата има по-голяма грешка, поради факта, че групата е по-нехомогенна.

Приложение на формулата за t-стойността

След като сме изчислили стандартните грешки на средните аритметични за двете групи заместваме във формулата:

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{{SE_{m_1}}^2 + {SE_{m_2}}^2}}$$

$$t = \frac{|65,00 - 81,83|}{\sqrt{0,75^2 + 1,29^2}}$$

$$t = \frac{|-16,83|}{\sqrt{0,56+1,66}} = \frac{16,83}{\sqrt{2,22}} = \frac{16,83}{1,49} = 11,3$$

Изчисляване на p стойността

За да определим вероятността разликата в двете групи да бъде случайна - p използваме изчислената t стойност. Не са необходими допълнителни формули, понеже не е необходима точната числена стойност на p. В тази стъпка само определяме отношението на вероятността p спрямо зададеното ниво на значимост (риск от грешка 1-ви род) α. За целта използваме таблица 7

таблица 7: t-разпределение

df	p = 0.1	**p= 0.05**	p = 0.01
30	t= 1.645	**t = 1.960**	t = 2.576

В таблицата:

- df е **степента на свобода**. Тя зависи от броя единици на наблюдения. Познаването на този индикатор не е необходимо за настоящият анализ - по правило се използват само стойностите отговарящи на $df > 30$;
- Всяка една колона представя гранично ниво на риска за грешка от първи род α. В медицината използваме лимитираща стойност от $\alpha < 0.05$, затова се интересуваме от стойностите във **втората колона**;
- По редове са представени стойностите на t критерия, отговорящи на съответните нива на значимост;
- Ако получим $t = 1,645$ вероятността за случайност е 10% (0.1). При $t = 1,96$ вероятността за случайност е 5% (0.1), а при $t = 2,576$ вероятността за случайност е 1% (0.01).

В нашия пример стойността $t = 11,3$. Въпреки, че тя липсва в таблица 7 можем да заключим, че вероятността за случайност е под 1% (0.01).

Статистически извод

Сега е момента да направим финалното заключение за нашето проучване.

- Разликата между двете групи е 16,83, а t критерия е 11,3.
- В таблицата установения t критерии е по-голям от 2.576, следователно и вероятността да наблюдаваме подобна разлика в двете групи е по-малка от от 0,01.

- Определената вероятност е под риска за грешка - p < α <0.05, което означава че отхвърляме H_0 и приемаме H_1 - групата лекувана с бета блокер има значително по-ниски стойности на сърдечния пулс спрямо пациентите лекувани с плацебо.

> ⚠ Важно
>
> Колкото е по-висока стойността на t критерия, толкова по-малко вероятно е да бъде наблюдаваме случайно. При стойност на t над 1,96, вероятността за случайност $p < 0.05$, което ни позволява да отхвърлям нулевата хипотеза

Важни зависимости

При равни други условия:

- По-голямата **абсолютната разлика** в средните аритметични води до по-голяма стойност на t критерия и до по-малка вероятност за случайност;
- По-високата **хомогенност** в групите води до по-малко стандартно отклонение, което е причина за по-малка стандартна грешка и по-високи стойности на t критерия;
- Увеличаването на **броя наблюдения** в извадката води до по-малко стандартно отклонение и по-малка стандартната грешка, тоест до по-високи стойности на t критерия.

T-тест за качествени признаци

T тестът може да се използва и за качествени признаци, когато искаме да сравним два относителни дяла. За целта се използва формулата:

$$t = \frac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{{SE_{p_1}}^2 + {SE_{p_2}}^2}}$$

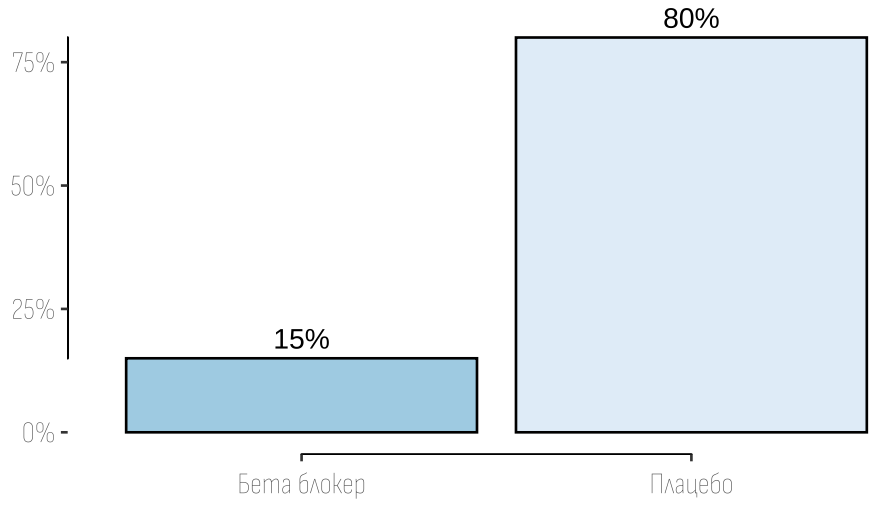
Където:

- $\hat{p}_1$ е относителният дял на група 1;
- $\hat{p}_2$ е относителният дял на група 2;
- SE_{p_1} е стандартната грешка на относителния дял в група 1;
- SE_{p_2} е стандартната грешка на относителния дял в група 2;
- $|\hat{p}_1 - \hat{p}_2|$ е абсолютната стойност на разликата между относителните дялове;

- $\sqrt{{SE_{p_1}}^2 + {SE_{p_2}}^2}$ е корен квадратен от сумата на квадратите на стандартните грешки на относителните дялове.

За да приложим формулата нека представим следния пример. Проследили сме пациентите ползващи бета блокер или плацебо за период от 5 години. Изследователите са записали броя на починалите от всяка една от двете групи.

Общо са починали 38 пациенти. В групата с бета блокер са починали 6-ма (15%), а в групата с плацебо 32-ма (80%). В фигура 2 прави впечатление, че пропорцията на починалите сред приемащите новото лекарство е в пъти по-малка от тази на контролната група.



фигура 2: Пропорция на починалите в двете групи пациенти

За да установим дали тази разлика е в резултат на случайност, отново използваме t-теста. В този пример обаче е необходима неговата модификация - вместо средните аритметични се използват относителните дялове (процент).

Дефиниране на нулева хипотеза H_0 и алтернативна хипотеза H_1

Нулева хипотеза Няма разлика между съотношението на починалите в двете групи пациенти. Дори и да наблюдаваме такава, тя е резултат на случайност и не съществува в генералната съвкупност.

Алтернативна хипотеза Има разлика между относителните дялове на починалите пациенти в двете групи. Тази разлика не е в резултат на случайност.

Определяне на риска за грешка (ниво на значимост)

Отново използваме нивото α под 0,05

Определяне на стандартната грешка

Използваме познатата от упражнение 2 формула:

$$SE_p = \sqrt{\frac{\hat{p} \cdot (100 - \hat{p})}{n}}$$

За пациентите ползващи бета блокер:

$$SE_{p1} = \sqrt{\frac{\hat{p_1} \cdot (100 - \hat{p_1})}{n}} = \sqrt{\frac{15 \cdot (100 - 15)}{40}} = \sqrt{\frac{15 \cdot 85}{40}} = \sqrt{\frac{1275}{40}} = \sqrt{31,88} = 5,6$$

За пациентите ползващи плацебо:

$$SE_{p2} = \sqrt{\frac{\hat{p_2} \cdot (100 - \hat{p_2})}{n}} = \sqrt{\frac{80 \cdot (100 - 80)}{40}} = \sqrt{\frac{80 \cdot 20}{40}} = \sqrt{\frac{1600}{40}} = \sqrt{40} = 6,32$$

Изчисляване на t-статистиката

Прилагаме формулата:

$$t = \frac{|\hat{p_1} - \hat{p_2}|}{\sqrt{{SE_{p1}}^2 + {SE_{p2}}^2}}$$

$$t = \frac{|15 - 80|}{\sqrt{5,6^2 + 6,32^2}} = \frac{|-65|}{\sqrt{31,36 + 40,45}} = \frac{65}{\sqrt{31,36 + 40,45}} = \frac{65}{\sqrt{71,81}} = \frac{65}{8,47} = 7,6$$

Определяме p стойността

След като получим стойността на t критерия проверяваме в коя посока се намира вероятността p в таблица 8

таблица 8: t-разпределение

df	p = 0.1	p= 0.05	p = 0.01
30	t= 1.645	t = 1.960	t = 2.576

Т критерият е по-висок от 2.576 (тоест е надясно от най-високата стойност), което означава, че вероятността p е по-малка от 0,01.

Заключение

При t = 7,6 вероятността за случайност p е по-малка от риска за грешка α < 0.05. Следователно, отхвърляме H_0 и приемаме H_1 - съществува статистически значима разлика в относителните дялове на починалите в двете извадки.

03-Упражнение 2023/2024

Домашна работа

ас.д-р Костадин Костадинов

Задачи

Задача 1

По случаен начин са подбрани две извадки с обем съответно n1 = 248 и n2 = 140 индивиди с кардиологични проблеми. В таблицата по-долу е регистриран пулсът на пациентите. Изчислете границите на 95% доверителен интервал за средните величини на популацията за двете извадки. Докажете дали средната аритметична е статистически различна спрямо групата пациенти, използвайки Т-тест.

Група 1 - пулс	Група 1 - брой	Група 2 - пулс	Група 2 - брой
60	4	80	10
65	18	85	15
70	25	90	25
75	30	95	30
80	50	100	28
85	48	105	16
90	36	120	2
95	28	125	8
100	9	130	6
Общо	248	Общо	140

Задача 2

Да се определи съществено ли е различието във виталния капацитет при индивиди от мъжки пол на 19 годишна възраст в две независими извадки с еднакъв брой случаи. В едната извадка са включени баскетболисти, а в другата младежи, трениращи плувен спорт. Да се приложи t тестът, като необходимите статистически параметри - средна аритметична и стандартно отклонение да се изчислят както при групирани данни.

Баскетболисти		**Плувци**	
Витален капацитет	Честота (f)	Витален капацитет	Честота (f)
2601-2800	12	3401-3600	13
2801-3000	21	3601-3800	26
3001-3200	35	3801-4000	35
3201-3400	62	4001-4200	58
3401-3600	33	4201-4400	30
3601-3800	19	4401-4600	20
3801-4000	15	4601-4800	15
общо:	197	общо:	197

Тестове

Задача 1

В рамките на клинично проучване пациенти със стомашен аденокарцином са разделени на 3 групи в зависимост от приложената терапия. Изследователският екип има за цел да сравни 4-седмичните изменения в абсолютния туморен размер спрямо изходните стойности, като тази величина не е нормално разпределена. Какъв тест за проверка на хипотеза следва да бъде приложен?

a) H-тест на Kruskal–Wallis
b) ANOVA
c) Тест на Friedman за свързани извадки
d) ANOVA с повторни измервания

Задача 2

В рамките на проспективно наблюдение на различни диетични режими за намаляване на теглото ученици със затлъстяване са разделени на 2 групи. 95% CI за средното намаление на теглото при 3-месечна диета А е между 4.5 и 7.2 кг. 95% CI за средното намаление на теглото при 3-месечна диета Б е между 2.2 и 4.2 кг. Какъв извод може да се направи?

a) Няма разлика в ефективността на двете диети.
b) Има статистически значима разлика в ефективността на двете диети.
c) Диета А е статистически значимо по-ефективна от диета Б.
d) Диета Б е статистически значимо по-ефективна от диета А.

Задача 3

В рамките на клинично проучване пациенти с множествена склероза са разделени на 2 групи в зависимост от приложената терапия. 95% CI за честотата на рецидиви при терапия А между 21 и 38%. 95% CI за честотата на рецидиви при терапия Б между 34 и 48%. Какъв извод може да се направи?

a) Няма разлика в ефективността на двете терапии.
b) Има статистически значима разлика в ефективността на двете терапии.
c) Терапия А е статистически значимо по-ефективна от терапия Б.
d) Терапия Б е статистически значимо по-ефективна от терапия А.

Задача 4

При равни други условия, ако намалим нивото на значимост от $\alpha = 0.05$ на $\alpha = 0.01$

a) Вероятността за грешка от I род ще нарастне, докато вероятността за грешка от II род ще намалее.
b) Вероятността за грешка от II род ще нарастне, докато вероятността за грешка от I род ще намалее.
c) Вероятността за грешки от I и от II род ще нарастне.
d) Вероятността за грешки от I и от II род ще намалее.

Задача 5

В проспективно проучване на пациенти с миастения гравис тежки миастенни кризи са наблюдавани при 23 (30.7%) от общо 75 пациенти, 95% CI за P (20.1%; 40.3%). Кое от следните твърдения е вярно?

a) В популацията от пациенти с миастения гравис честотата на тежки миастенни кризи най-вероятно е по-голяма от 30.7%.
b) Произволен пациент с миастения гравис има 30.7% шанс да развие тежки миастенни кризи.
c) Има 95% вероятност честотата на тежки миастенни кризи в изследваната извадка да бъде между 20.1% и 40.3%.
d) В популацията от пациенти с миастения гравис честотата на тежки миастенни кризи е малко вероятно да бъде по-малка от 20.1%.

Задача 6

В рамките на клинично проучване пациенти с белодробен карцином са разделени на 2 групи в зависимост от приложената терапия. Резултатите са анализирани с помощта на t-тест за сравнение на 2 независими извадки. Пациентите, лекувани с експериментална терапия, преживяват средно 10.1 месеца, докато лекуваните с химиотерапия преживяват средно 8.7 месеца (p = 0.08). Кое от следните твърдения е вярно?

a) Приемаме нулевата хипотеза, че има статистически значима разлика в средната преживяемост при двата вида терапия.
b) t-тестът за сравнение на 2 независими извадки е подходящ за сравнение на средната преживямост, тъй като в случая средните стойности за този показател са приблизително равни в двете групи.
c) Тъй като резултатът не е статистически значим, може да преминем към анализ на данните с t-тест за две свързани извадки.
d) Подходяща алтернативна хипотеза в случая е, че двете групи пациенти имат различна преживяемост в зависимост от вида на лечението.

Задача 7

В проучване на родилки с множествена склероза средното тегло на новородените е 3 100 г при майките, които са в ремисия от повече от 1 година, и 2 850 г при майките, които са в ремисия по-малко от 1 година (p = 0.06). Резултатите са анализирани с помощта на t-тест за сравнение на 2 независими извадки. Кое от следните твърдения е вярно?

a) Полученият резултат е статистически значим при α = 0.05, което означава, че нулевата хипотеза може да бъде отхвърлена.
b) Тест на McNemar също може да се използва за тест на хипотеза в този случай.
c) Теглото на тези новородени най-вероятно не е нормално разпределена величина.
d) Нулевата хипотеза гласи в случая, че средното тегло на новородените е еднакво за двете групи родилки.

Задача 8

Кое от следните твърдения е вярно?

a) Случайната грешка се дължи на обстоятелството, че се наблюдава част от генералната съвкупност.
b) Систематичната грешка се дължи на получена недостоверна информация.
c) Грешка на оценката се получава при извадкови изследвания и включва 2 вида грешки – стохастична и нестохастична.
d) Всички отговори са верни.

Задача 9

Кое от следните твърдения е вярно?

a) t-тест за 2 независими извадки е приложим единствено, когато изследваната променливата е нормално разпределена.
b) t-тест за 2 независими извадки е подходящ, когато искаме да сравним 2 променливи, всяка от които се измерва еднократно за всеки индивид от извадката.
c) t-тест за 2 независими извадки е подходящ, когато искаме да сравним 1 променлива, измерена за индивиди преди началото на клинично проучване, с референта стойност от терапевтичен стандарт.
d) Всички отговори са грешни.

Задача 10

При равни други условия, най-малък необходим брой единици на наблюдение ще бъде необходим при: a) Разнородна популация и голям по размер клиничен ефект. b) Еднородна популация и голям по размер клиничен ефект. c) Разнородна популация и малък по размер клиничен ефект. d) Еднородна популация и малък по размер клиничен ефект.

Отворени въпроси

Задача 1

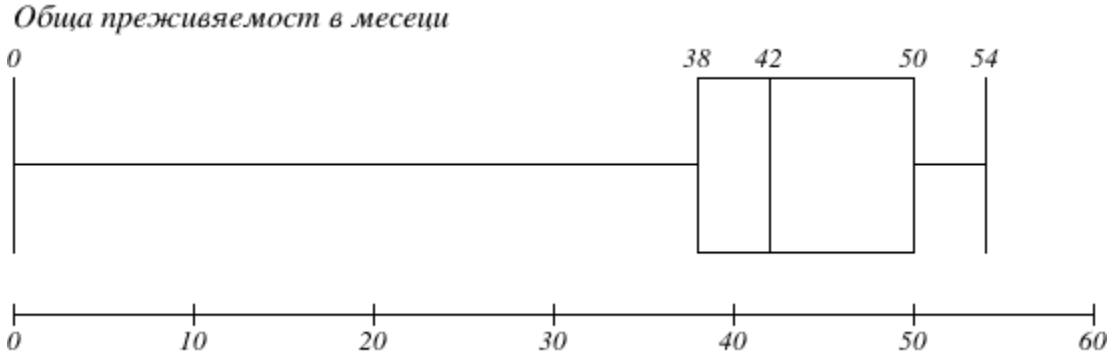
Каква е стойността на интерквартилния размах на променливата „обща преживяемост в месеци" (представена на боксплот диаграмата по-долу)? Какво по вид е разпределението на тази променлива? Има ли и каква е стойността на най-екстремните аутлайъри в този случай?

Задача 2

Може ли стандартното отклонение да бъде равно на нула? Обяснете защо.

Задача 3

Обемът на извадката оказва ли влияние върху вероятността за грешки от II? А върху вероятността за грешки от I род? Обяснете защо.



фигура 1: Задача 1

Задача 4

Мощността на научноизследователския дизайн при проверка на хипотеза зависи от 4 основни фактора. Дайте пример за комбинация от тях, която ще постигне дизайн с възможно най-голяма мощност.

Задача 5

Извадка от 2 000 новородени е включена в проспективно изследване. Средното тегло на новородените при раждане е 3 500 г със стандартно отклонение 400 г. Теглото при раждане е нормално разпределена величина. Колко от включените за наблюдение новородени имат тегло, което е по-голямо от 4 300 г или по-малко от 3 100 г?

Задача 6

Проучване на нивото на физическа активност (измервано като брой часове тренировка седмично) при пушачи и непушачи достига до t = 2,96. Какво заключение може да се направи въз основа на този резултат?

04 & 05 - Упражнение 2023/2024

Изследване на връзки. Корелация. Асоциация

ас.д-р Костадин 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 свръзка**. С други думи, резултатите от статистическия анализ сами по себе си **не могат** да се разглеждат като изчерпателно доказателство на тезата за причинната обусловеност[1].

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 познанието за явленията [2].
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 явленията [3].

[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-улесненитe 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, заболеваемост, тютюнопушене, диагноза) се изследва с помощта на теста за **асоциация**.

Статистическата връзка се изследва в **обема** 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:

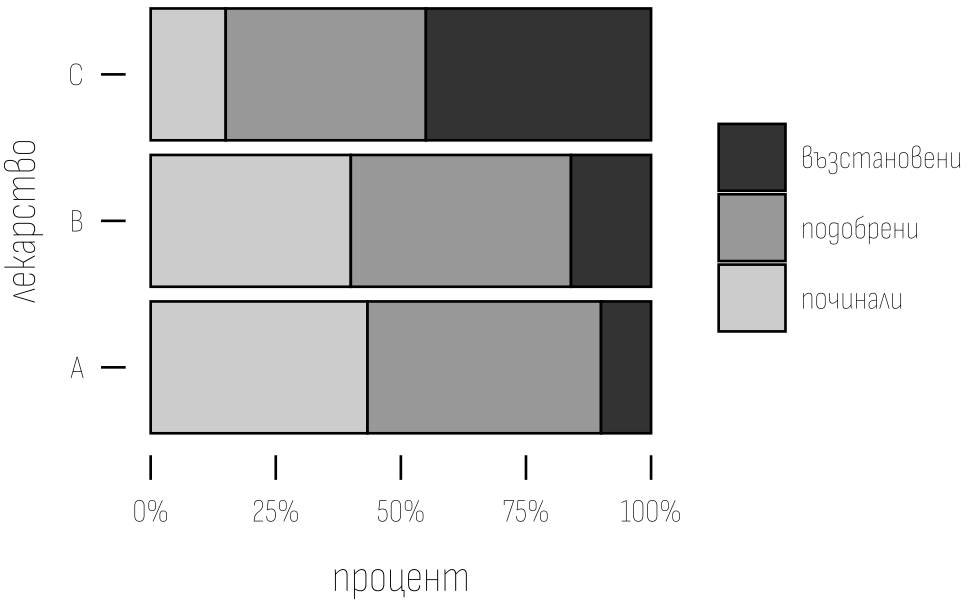
1. Дефиниране на нулева и алтернативна хипотеза;
2. Определяне на ниво на грешка;
3. Изчисляване на теоретични честоти;
4. Изчисляване на стойността на χ^2;
5. Определяне на p-стойност;
6. Заключение.

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 "A", "B" и "C". В изследването са включени 500 пациенти, от които лекувани с "A" са 150, с "B" - 250, а с "C" - 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-ство	починали	подобрение	оздравели	$\sum_{rows}$
A	$f_o a$ **65**	$f_o d$ **70**	$f_o g$ **15**	150
B	$f_o b$ **100**	$f_o e$ **110**	$f_o h$ **40**	250
C	$f_o c$ **15**	$f_o f$ **40**	$f_o i$ **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 лечението [5].

Алтернативната хипотеза е противоположна на нулевата и твърди, че същества 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:

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

Където:

- $f_t cell$ 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":

$$f_t a = \frac{180 \cdot 150}{500} = 54$$

Теоретичната честота $f_t a$ означава, че ако липсва връзка между лечението и резултата, починалите пациентите 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-во	починали	подобрение	оздравели	$\sum_{rows}$
A	$f_t a$ *54*	$f_t d$ *66*	$f_t g$ *30*	150
B	$f_t b$ *90*	$f_t e$ *110*	$f_t h$ *50*	250
C	$f_t c$ *36*	$f_t f$ *44*	$f_t i$ *20*	100
$\sum_{columns}$	180	220	100	*500*

За всички останали клетки

$$f_t b = \frac{180 \cdot 250}{500} = 90$$

$$f_t c = \frac{180 \cdot 100}{500} = 36$$

$$f_t d = \frac{220 \cdot 150}{500} = 66$$

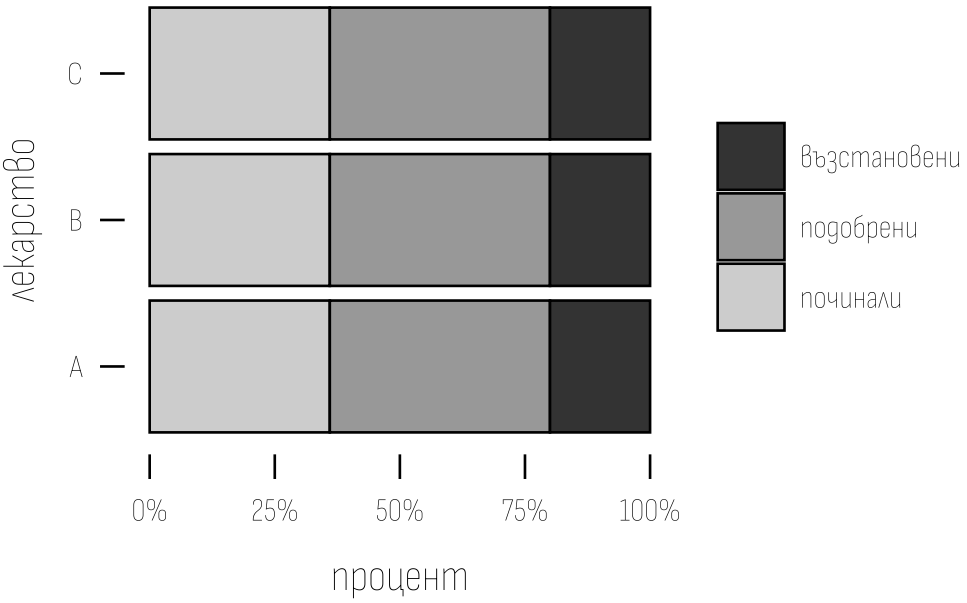
$$f_t e = \frac{220 \cdot 250}{500} = 110$$

$$f_t f = \frac{220 \cdot 100}{500} = 44$$

$$f_t g = \frac{100 \cdot 150}{500} = 30$$

$$f_t h = \frac{100 \cdot 250}{500} = 50$$

$$f_t i = \frac{100 \cdot 100}{500} = 20$$



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 p-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 стъпки трябва да определялият "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:

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

Във формулата:

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

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.

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

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

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

Заключение

$\chi^2 = 57$ 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 (таблици 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 -лата, толкова **по-силна** е наблюдаваната статистическа връзка. Съществуват множество методики и видове корелационнни коефициенти, въпреки това целта на всички тях е:

- Да оцени **силата на статистическата връзка**
- Да се оцени **посоката** на връзката [7];
- Да се определи **формата** на връзката [8];

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 връзката между тях е **линейна**. Интерпретацията на коефициента на Пирсън е представена в таблица 5.

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;

[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**.

таблица 5: ρ интерпретация

ρ	сила и посока на връзката
1	"перфектна" *права*
0,67 до 1	силна *права*
0, 34 до 0, 66	умерена *права*
0 до 0,33	слаба *права*
0	**липсва връзка**
0 до - 0,33	слаба *обратна*
- 0,34 до -0,66	умерена *обратна*
- 0,67 до - 1	силна *обратна*
-1	"перфектна" *обратна*

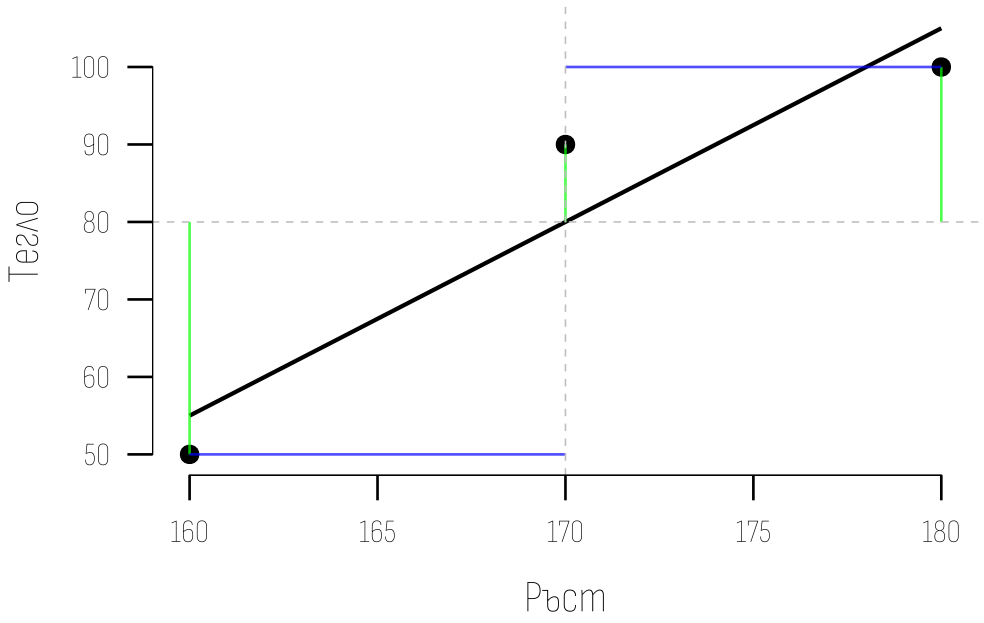
- $x_i - \bar{x}$ е отклонението на всяка една точка от средната аритметична за променливата x;
- $y_i - \bar{y}$ е отклонението на всяка една точка от средната аритметична за променливата y;
- $\sum_{i=1}^{n}(x_i - \bar{x})^2$ е сумата на квадратите на отклоненията за променливата x;
- $\sum_{i=1}^{n}(y_i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

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

За втория студент (Иван):

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

За третия студент (Чавдар):

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

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

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

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 хипотеза [9]

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

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 че в действителност има корелация и тя не е равна на 0-ла.

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

04-Упражнение 2023/2024

Домашна работа

ас.д-р Костадин Костадинов

Изчислителни задачи

1

Като използвате **непараметричен тест хи квадрат** проверете съществува ли асоциация между продължителността трудовия стаж при миньорите и заболеваемостта от вибрационна болест?

трудов стаж	с вибрационна болест	здрави	Общо
до 5г.	32	127	159
6-15г.	56	268	324
16-25г.	23	48	71
над 25г.	41	34	75
Общо	152	477	629

2

При проучване на влиянието на вида използван дюшек при получаване на декубитусни рани, са получени следните резултати:

- При *low air-loss* дюшек при **7 от 56** изследвани случаи имат декубитусни рани
- При *воден дюшек* при **13 от 71** изследвани лица
- При дюшек *на вълни* с декубитусни рани са **25 от 72** пациенти

Определете дали видът на използвания дюшек е фактор за появата на декубитосни рани?

3

При проучване на нов анестетик е отчетено времето от подаването му във венозната система до реакцията на пациента и съответната ефективна доза, определена в ug/kg.

доза	време за въздействие
1,3	25
1,9	13
1,5	19
1,1	23
1	26
1,3	21
2,1	11
1,7	30
1,6	14
0,9	24
общо: 14,4	общо: 206

Първо:

Изчислете 95% -вия интервал на доверителност на ефективна доза в генералната съвкупност.

Второ:

Изчислете 95% -вия интервал на доверителност за времето до ефект в генералната съвкупност.

Трето:

Докажете дали съществува връзка между явленията **доза** и **време за реакция**.

Тестови задачи

1. Кой вид диаграма е най-подходящ за представяне на връзката между 2 количествени променливи?

a) Комбинирана диаграма на разсейването (scatterplot)
b) Стълбовидна диаграма (bar chart)
c) Хистограма

d) Кръгова диаграма (pie chart)

2. Корелационен коефициент r = -0.10 означава:

a) Силна обратна причинно-следствена връзка.
b) Умерена обратна причинно-следствена връзка.
c) Слаба обратна причинно-следствена връзка.
d) Всички отговори са грешни.

3. Кое от следните твърдения е вярно?

a) Интерполацията прогнозира извън област от вече известни стойности, които са послужили за конструиране на регресионен модел.
b) При равни други условия, екстраполацията постига по-голяма валидност на получената оценка в сравнение с интерполацията.
c) Екстраполацията поставя условие наблюдаваната зависимост да се запази извън областта от вече известните стойности.
d) Графичният образ на еднофакторните нелинейни модели са различни прави в равнината.

4. Фи коефициентът за корелация на Пирсън се използва при анализ на връзката между:

a) 2 количествени променливи.
b) 2 дихотомни променливи.
c) 1 количествена и 1 качествена променливи.
d) 1 количествена и 1 дихотомна променливи.

5. Корелационен коефициент r = -1.2 означава:

a) Силна обратна зависимост.
b) Умерена обратна зависимост.
c) Слаба обратна зависимост.
d) Всички отговори са грешни.

Отворени въпроси

1. Проучване на нивото на затлъстяване (дефинирано като „предзатлъстяване" „затлъстяване I степен" „затлъстяване II степен" и „затлъстяване III степен") при пушачи и непушачи достига до резултат $\chi2 = 9.488$ / df = 4. На колко е равна p-стойността в този случай и какво означава това? Какъв извод може да се направи въз основа на този резултат?

2. Проучване на нивото на затлъстяване при пушачи и непушачи достига до резултат $\chi2 = 15.507$ / df = 8. На колко е равна p-стойността в този случай и какво означава това? Какъв извод може да се направи въз основа на този резултат?

06-Упражнение 2023/2024

Домашна работа

ас.д-р Костадин Костадинов

Изчислителни задачи

Задача 1

Данните регистрирани в таблицата по-долу са на пациенти, участници в проучване на коронарната болест на сърцето. Изградете регресионен модел с който да предскажете стойността на холестерола в кръвта на пациентите спрямо теглото им.

Тегло (кг)	Холестерол (мг/дл)
84	135
107	222
94	244
83	252
131	269
72	286
87	294
88	302
77	311
90	311
79	312
82	353
70	403
88	403

Задача 2

При проучване на нов анестетик е отчетено времето от подаването на медикамента до реакцията на пациента и съответната ефективна доза, определена в мкг/кг. Данните са показани в таблицата по-долу. Изградете регресионен модел с който да времето за реакция в зависимост от дозата на медикамента.

Доза (мкг/кг)	Време (сек)
1,3	25
1.9	13
1,5	19
1,1	23
1	26
1,3	21
2,1	11
1,7	30
1,6	14
0,9	24

Затворен тип въпроси

1. Множествената (многофакторна) линейна регресия моделира:

a) Резултат за 1 независима променлива въз основа на данни за 1 зависима променлива.
b) Резултат за 1 зависима променлива въз основа на данни за 1 независима променлива.
c) Резултат за 1 независима променлива въз основа на данни за повече от 1 зависими променливи.
d) Резултат за 1 зависима променлива въз основа на данни за повече от 1 независими променливи.

2. Големият положителен наклон на регресионната крива означава, че:

a) корелацията е силна и положителна;
b) вертикалното изместване е голямо и положително;
c) стойността на Y ще се увеличи поне с няколко единици, когато стойността на X се увеличи с една единица;
d) всички гореизброени са верни.

3. Ако коефициентът на детерминация е положителен, то простата (единична) линейна регресия:

a) Има положителен (нарастващ) наклон.
b) Има отрицателен (намаляващ) наклон.
c) Може да има или положителен, или отрицателен наклон.
d) Зависимата променлива Y е отрицателна.

4. Каква е целта на простата (единична) линейна регресия?

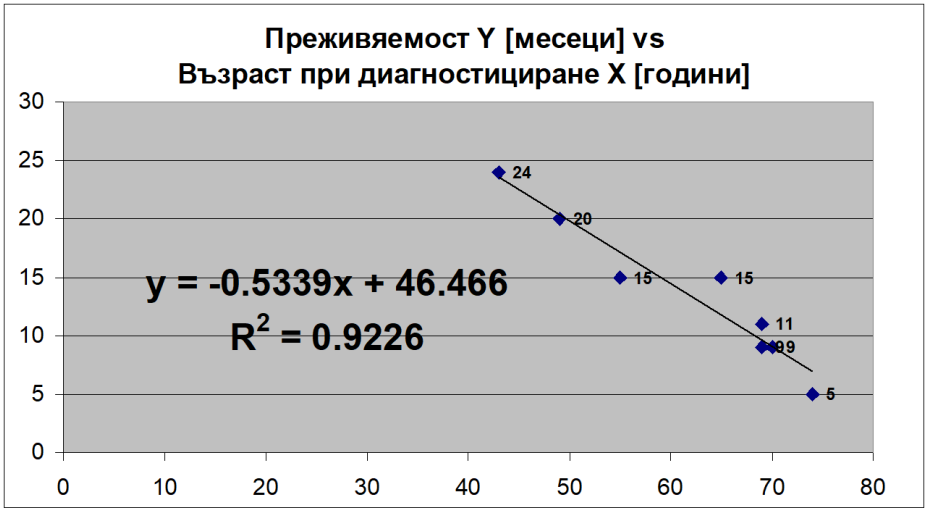
a) Прогнозиране на зависима променлива въз основа на данни за една независима променлива.
b) Прогнозиране на независима променлива въз основа на данни за една зависима променлива.
c) Прогнозиране на зависима променлива въз основа на данни за повече от една независими променливи.
d) Прогнозиране на независима променлива въз основа на данни за повече от една зависими променливи.

5. Регресионен коефициент b = 0.85 означава:

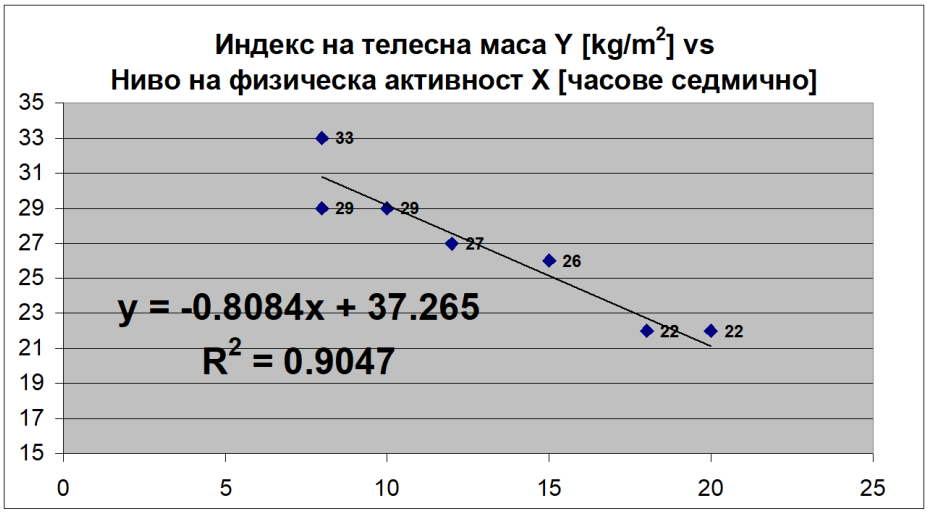
a) Положителен наклон на правата на най-добро съответствие
b) Отрицателен наклон на правата на най-добро съответствие
c) Наклонът на правата на най-добро съответствие може да бъде както положителен, така и отрицателен
d) Всички отговори са грешни

Отворен тип въпроси

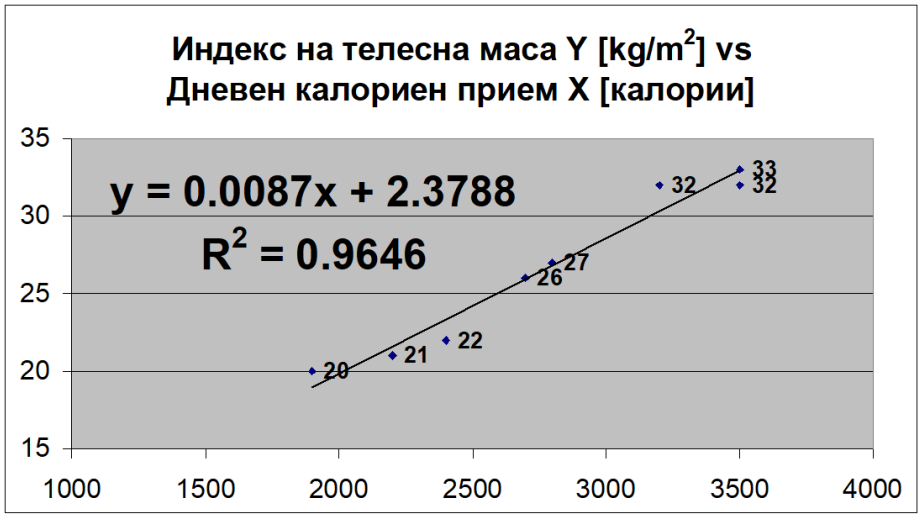
Какъв извод може да се направи въз основа на регресионния модел, представен на графиката по-долу? Как интерпретираме коефициента на детерминация в този случай? Каква е стойността на коефициента на корелация?



фигура 1: Вариант 1



фигура 2: Вариант 2



фигура 3: Вариант 3

06 - Упражнение 2023/2024

Регресионен анализ. Линейна регресия

ас.д-р Костадин Костадинов

Същност на регресионния модел

Когато две променливи са в корелация помежду си (било то положителна или отрицателна), познаването на стойността на едната променлива позволява да **предскаже**, сравнително точно, стойността на другата. Това позволява да се правят полезни прогнози. Подобни модели са изключително важни в епидемиологията и клиничната медицина, а най-често използваният метод за тяхното създаване е **линейната регресия**.

Променливата, чиито стойности се предсказват в регресионния модел се нарича зависима (резултативна), докато променливата, която се използва за предсказване и построяването на модела се нарича независима (факторна). В своята същност, регресионния анализ достига до "уравнение" (математическа функция) чрез което в аналитичен вид се представя връзката между определен набор от фактори и резултативното явление. Регресионното уравнение се представя като:

$$y = a + \beta \cdot x$$

Нека приемем за зависима променлива y - тегло на студентите от 2-ри курс, а за независима x - калорийния им прием. Чрез линейната регресия връзката между тези променливи може да се моделира, така че при известни стойности за калорийния прием, да предскажем теглото на студента. Регресионният модел следва формулата: $y = a + \beta \cdot x$, в която:

- Коефициентът a е **константа** (intercept).

Стойността коефициента показва, какво е теглото, когато калорийният прием е нула. С други думи, a посочва началната точка на взаимовръзката между тегло и калорийния прием и на практика, няма практическо значение в интерпретацията на модела, тя служи единствено за математическото моделиране.

- β е **регресионният коефициент**.

Това е число, което показва **силата и посоката** на връзката между предиктора (калорийния прием) и резултативната променлива (теглото). Математически, за установяване на стойността на y, трябва да умножи регресионния коефициент β с известната стойност калорийния прием x. Отрицателните стойности на β означават негативна (обратна) корелационна връзка, а позитивните стойности - позитивна (права).

Регресионно моделиране

Създаването на линеен регресионен модел протича в следните стъпки:

1. Избор на зависима и независима променлива
2. Изчисляване на регресионния коефициент β
    1. Изчисляване на сумата на стойностите на променливите x и y, както и сумата на всички произведения между на x по y за всяко едно наблюдение;
    2. Сбор на повдигнатите на квадрат стойности на променливата x и умножение на получената сумата с броя на наблюденията;
3. Изчисляване на константата a
    1. Изчисляване на средната стойност на променливата y;
    2. Изчисляване на средната стойност на променливата x;
4. Интерпретация на модела.

За да се представят етапите в тяхната последователност, се използва следния пример

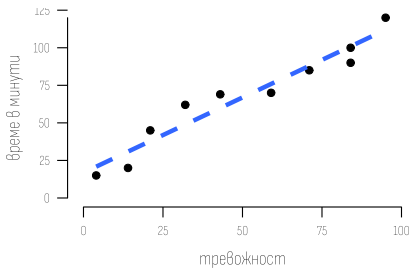
> Пример
>
> Проучване измерва степента на тревожност сред 10 студенти в специалност "медицина", посредством въпросник със скала от 0 (най-ниска степен на тревожност) до 100 (най-висока степен на тревожност). Изследователският екип записва резултата от теста, както и времето (в минути) прекарано в четене по статистика в предходния ден за всеки един участник. Целта на изследването е да се построи **линеен регресионен модел** за предсказване на тревожността на студентите y въз основа на времето прекарано в четене x.

Избор на зависима и независима променлива

Изходните данни са представени в таблица 1 и на фигура 1. Променливата `тревожност` е зависима, това означава, че изградения модел ще се опитва да *предскаже* стойността на тревожността въз основа на променливата `учене по статистика`, която е независима или факторна.

таблица 1: Изходни данни за регресионен модел

y Тревожност (0 до 100)	x учене по статистика (минути)
21	45
84	90
14	20
95	120
84	100
32	62
59	70
43	69
71	85
4	15



фигура 1: Връзка между тревожност и време за учене по статистика

Изчисляване на регресионния коефициент бета

За изчисление на този коефициент се ползва формулата:

$$\beta = \frac{(n \cdot \sum(X \cdot Y)) - (\sum X \cdot \sum Y)}{(n \cdot \sum X^2)) - (\sum X)^2}$$

Изчисляване на сумата на стойностите на променливите x и y, както и сумата на всички произведения между на x по y за всяко едно наблюдение;

В тази стъпка е необходимо да установи сборът на всички наблюдения за двете променливи. След това е необходимо да се установи сумата $\sum(X.Y)$ от произведенията на стойностите на променливите x по y - за за всеки един студент. Резултатите са представени таблично в таблица 2.

В числител:

- n е броят на наблюденията;
- $\sum X$ е сумата от стойностите на независимата променлива x;
- $\sum Y$ е сумата от стойностите на зависимата променлива y;
- $\sum(X.Y)$ е сумата от всички произведения на стойностите на x по стойностите на y за всяко едно наблюдение.

В знаменателя:

- n е броят на наблюденията;
- $\sum X^2$ е сумата от квадратите на стойностите на независимата променлива x;
- $(\sum X)^2$ е квадратът на сумата от стойностите на независимата променлива x.

таблица 2: Първа стъпка за изчисляване на регресионния коефициент

y Тревожност	x учене	$x \cdot y$
21	45	945
84	90	7560
14	20	280
95	120	11400
84	100	8400
32	62	1984
59	70	4130
43	69	2967
71	85	6035
4	15	60
$\sum Y = 507$	$\sum X = 676$	$\sum (X.Y) = 43761$

В резултат от тази стъпка се установява числителят в формулата за изчисляване на регресионния коефициент β

$$(n \cdot \sum (X \cdot Y)) - (\sum X \cdot \sum Y) = 10 \cdot 43761 - 676 \cdot 507 = 437610 - 342732 = 94878$$

Сбор на квадратите за променливата x

В тази стъпка се повдигат на квадрат всички стойности на променливата x, след което се сумират. Резултатът е представен в таблица 3.

таблица 3: Втора стъпка за изчисляване на регресионния коефициент

x учене по статистика (минути)	x^2
45	2025
90	8100
20	400
120	14400
100	10000
62	3844
70	4900
69	4761
85	7225
15	225
$\sum x = 676$	$\sum x^2 = 55880$

В края на втора стъпка получаваме знаменателя в формулата за изчисляване на регресионния коефициент β

$$(n \cdot \sum X^2) - (\sum X)^2 = 10 \cdot 55880 - 676^2 = 558800 - 456976 = 101824$$

Изчисляване на регресионния коефициент β

След като сме установили числителя и знаменателя заместваме във формулата:

$$\beta = \frac{(n \cdot \sum X \cdot Y) - \sum X \cdot \sum Y}{n \cdot \sum X^2 - (\sum X)^2}$$

$$\beta = \frac{94878}{101824} = 0,93$$

Изчисляване на константата a

За изчисляване на константата използваме формулата:

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

Където:

- $\bar{y}$ е средната стойност на зависимата променлива y;
- $\bar{x}$ е средната стойност на независимата променлива x.
- β е регресионният коефициент.
- a е константата.

Изчисляване на средната стойност на зависимата променлива y

За да изчислим средната стойност на зависимата променлива y ползваме формулата за средна аритметична:

$$\bar{y} = \frac{\sum Y}{n}$$

$$\bar{y} = \frac{507}{10} = 50,7$$

Изчисляване на средната стойност на независимата променлива x

Отново прилагаме формулата:

$$\bar{x} = \frac{\sum X}{n}$$

$$\bar{x} = \frac{676}{10} = 67,6$$

Изчисляване на константата a

Заместваме в израза за константата:

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

$$a = 50,7 - (0,93 \cdot 67,6)$$
$$a = -12.2$$

Интерпретация:

1. **Регресионният коефициент** е позитивно число - това означава, че увеличаването на времето в четене по статистика, се свързва с увеличение на тревожността. Връзката е сравнително силна, като увеличение с една минута четене води до увеличение с 0,93 точки в теста за тревожност;
2. **Константата "а"** има значение единствено като *"начална точка"*. Това число показва стойността на зависимата променлива (тревожност), ако фактора (време прекарано в четене) е равен на 0.

Математически моделът се представя като:

$$y = a + \beta \cdot x = -12,2 + 0,93 \cdot x$$

Регресионно прогнозиране (използване на модела за предсказване)

Да приемем, че нов студент се премести в групата. Знаейки само колко време чете по статистика е необходимо да преценим, какъв би бил резултатът му от теста за тревожност. Да приемем, че студентът е споделил, че прекарва в четене 30 мин на ден. Прилагаме модела [1]:

$$y = a + \beta \cdot x = -12,2 + 0,93 \cdot x$$

За новия студент $x = 30$ минути. Заместваме в уравнението:

$$y = -12,2 + 0,93 \cdot 30 = 16$$

[1] Много компютърни програми използват методи подобни на регресията за целите на маркетинга. Някои интернет страници запаметяват последните ви търсения и в основа на данни за вас (независими променливи или фактори) ви "предлагат" определени продукти, които за които има по-голяма вероятност да се интересувате (зависима променлива). В медицината регресия се използва наравно с медицинската слушалка. В кардиологията например рискът от инсулт се изчислява, използвайки именно регресионен модел наречен *CHADVAS score*

Оценка на регресионния модел

Всеки регресионен модел се оценява спрямо **способността му да предсказва правилно**. Тази оценка се извършва чрез показателя коефициент на определеност (**детерминирания**)- R^2. Този коефициент:

1. Има стойности между 0 и 1 (0 и 100%);
2. Определя силата на модела - способността му да предсказва правилно;
3. Показва каква част от вариацията на зависимата променлива се обяснява от вариацията на независимата променлива;
4. Позволява да се оцени коефициентът на корелация, чрез формулата:

$$|r| = \sqrt{R^2}$$

В посоченият по-горе пример, ако приемем, че коефициентът на детерминирания е 0,93, това означава, че **93%** от вариацията в тревожността в групата се обяснява следствие на вариацията в времето за четене по статистика. Двете променливи са в права (позитивна) корелационна връзка, чиито коефициент е:

$$|r| = \sqrt{R^2} = \sqrt{0},93 = \pm 0,946$$

Общо правило за интерпретация на регресията:

1. Регресионният коефициент β, показва, как се променя резултатът (зависимата променлива) при **една единица повишаване на фактора** (предиктора).
2. Позитивен регресионен коефициент означава позитивна корелационна връзка между фактора и резултата. Обратно, негативен регресионен коефициент означава обратна (негативна) връзка между фактора и резултата.
3. R^2 представя какъв дял от вариацията на резултативната променлива се дължи на изследвания фактор.
4. Когато коренуваме стойността на R^2 получаваме **абсолютната стойност на корелационния коефициент** $\sqrt{R^2} = |r|$.
5. Знакът + или - пред коефициентът на корелация **се определя** от знака пред регресионния коефициент β.

Още малко за регресията

> **Множествена регресия**
>
> Представения до момента регресионен модел съдържа само една "обясняваща" променлива. Затова такива модели се означават като "еднофакторни".В действителност, за да обясним една променлива, често се използват множество предиктори (независими променливи). Всеки един от тях има собствен регресионен коефициент. В тези случай често се наблюдава статистическия феномен "колинеарност". При него две или повече независими са в **силна корелационна връзка помежду си**, при което тяхната способност да предсказват зависима променлива намалява.