МЕДИЦИНСКИ УНИВЕРСИТЕТ — ПЛОВДИВ
ФАКУЛТЕТ ПО ОБЩЕСТВЕНО ЗДРАВЕ
Катедра „Социална медицина и обществено здраве“

МЕДИЦИНСКА СТАТИСТИКА

Практически наръчник за упражнения

$\bar{X}$ SD $95\%\,CI$ p χ^2 r $\hat{Y}$

седем упражнения • решени примери • самоподготовка

гл. ас. д-р Костадин Костадинов, д.м., МЗМ

Учебно помагало за студенти от специалностите
„Медицина“ и „Дентална медицина“, II курс

Учебна 2026/2027 година
Пловдив

Медицинска статистика. Практически наръчник за упражнения

Автор: гл. ас. д-р Костадин Костадинов, д.м., МЗМ

Учебно помагало за практическите упражнения по медицинска статистика. Наръчникът следва програмата на курса. Той е предназначен за подготовка преди упражнение, за работа в час и за самоподготовка преди колоквиума и семестриалния изпит.

Всички изчисления, таблици и графики в наръчника са получени възпроизводимо със статистическата среда R и пакета kkstatfun. Данните в примерите са учебни: те са построени така, че да илюстрират метода, и не описват реални пациенти.

Забележки, открити грешки и предложения са добре дошли.

Съдържание

Предговор

Този наръчник е кратък учебен пътеводител за студентите от II курс по медицина и дентална медицина, които изучават **медицинска статистика**. Той следва програмата на седемте практически упражнения и съчетава обяснителен текст, решени примери и задачи за самоподготовка. Целта му е да свърже лекционните понятия с изчисленията и интерпретациите, необходими по време на упражненията, колоквиума и изпита.

Защо изобщо учим статистика?

Статистиката предоставя методите, чрез които данните се превръщат в информация, а информацията подпомага решенията. За лекаря това има три непосредствени приложения.

Първо, тя подпомага **информирания избор** в ежедневната практика. Когато преценявате дали да приложите ново лечение, трябва да разберете колко голям е наблюдаваният ефект, колко несигурна е оценката и доколко резултатът е приложим за конкретния пациент.

Второ, тя показва **как се изгражда научно доказателство**. Клиничните проучвания, препоръките в ръководствата и оценката на лекарствата използват статистически анализ, за да разграничат устойчивия резултат от случайното колебание.

Трето, тя позволява медицинската литература да се чете **критично**. Умението да разпознаете неподходящ метод, подвеждаща графика или прекалено силно заключение е по-ценно от запомнянето на отделен числов резултат.

> **Какво няма да научите в този курс**
>
> Съвременната статистика е основата на „науката за данните“ и на системите с изкуствен интелект. В този курс няма да програмирате и няма да строите сложни модели. Целта е по-скромна и по-важна: да разберете **логиката** на статистическия анализ и **езика**, на който той се говори — така че да можете да разчетете резултат, да прецените дали му вярвате и да обясните на пациент какво означава той.

Как е устроен наръчникът

Наръчникът съдържа седем глави — по една за всяко практическо упражнение, — глава за колоквиума и четири приложения.

Таблица 1: Съдържание на практическите упражнения

Гл.	Упражнение	Основен метод
1	Абсолютни и относителни величини	пряка стандартизация
2	Дескриптивна статистика	средна аритметична, медиана, SD, IQR
3	Инферентна статистика	средна грешка, доверителен интервал
4	Проверка на хипотези	Z-тест и t-тест
5	Връзка между качествени признаци	χ^2-тест, φ, V на Cramér
6	Корелационен анализ	коефициент на Pearson r
7	Регресионен анализ	проста линейна регресия

Всяка глава е построена по един и същ начин:

- **Цел на упражнението** — какво трябва да умеете в края на часа;
- **Какво трябва да знаете отпреди** — връзка с предходните упражнения;
- **Теория накратко** — определения и формули, всеки символ обяснен;
- **Решен пример** — стъпка по стъпка, с всички междинни числа;
- **Чести грешки** — какво се бърка най-често и защо;
- **Задачи за самостоятелна работа** — с кратки отговори в Секция Г;
- **Кратък тест** — въпроси с един верен отговор за проверка на разбирането.

Как да работите с него

Прочетете теоретичната част **преди** упражнението. По време на часа следвайте стъпките на решените примери и записвайте междинните изчисления. След упражнението решете задачите и краткия тест, без да поглеждате отговорите; сверете резултатите едва след като сте формулирали и словесен извод.

> **!** Три правила, които важат за целия курс
>
> 1. **Пишете стъпките, не само отговора.** На колоквиума и на изпита се оценява ходът на решението. Отговор без стъпки не носи точки.
> 2. **Всяко число има мерна единица.** „Средната преживяемост е 8“ не е отговор; „8 месеца“ е.
> 3. **Всяко заключение се формулира с думи.** Числото $p < 0{,}01$ не е извод. Изводът е изречението, което следва от него.

Означения

В целия наръчник се използват едни и същи символи. Латинските букви означават показатели, изчислени **в извадката** (статистики), а гръцките — съответните неизвестни показатели **в генералната съвкупност** (параметри).

Таблица 2: Основни означения

Символ	Значение
n	обем на извадката (брой наблюдения)
$\bar{X}$	извадкова средна аритметична (статистика)
μ	популационна средна аритметична (параметър)
S (или SD)	извадково стандартно отклонение
$S_{\bar{X}}$ (или SE)	средна грешка на средната аритметична
$\hat{p}$	извадков относителен дял, %
P	популационен относителен дял, %
f	честота (брой наблюдения в група или интервал)
α	ниво на значимост (допустим риск за грешка от I род)
p	p-стойност
df	степени на свобода

За десетичния знак

В българската научна практика за десетичен знак се използва **запетая** ($8{,}33$), а в англоезичната литература и в компютърните програми — **точка** (8.33). В наръчника е използвана запетая. И двата записа са верни; важното е да не се смесват в едно решение.

Използвана и препоръчана литература

- Сепетлиев, Д. (1980). *Медицинска статистика*. Трето преработено и допълнено издание. Медицина и физкултура.
- Димитров, И. (1996). *Медицинска статистика*. Второ преработено и допълнено издание. Пигмалион.
- Искров, Г. (2025). *Медицинска статистика*. УИ „Паисий Хилендарски“.
- Kirkwood, B., Sterne, J. (2003). *Essential Medical Statistics*. 2nd ed. Blackwell.
- Diez, D., Çetinkaya-Rundel, M., Barr, C. (2019). *OpenIntro Statistics*. 4th ed.

Изчисленията, таблиците и графиките са направени със статистическата среда **R** и пакета **kkstatfun**. Кодът не е показан — той не е част от изпитния материал, но е достъпен при поискване за студенти, които искат да го разгледат.

1 Абсолютни и относителни величини. Пряка стандартизация

Цел на упражнението

След това упражнение трябва да умеете:

1. да различавате абсолютна от относителна величина;
2. да разпознавате **екстензивен** от **интензивен** показател и да изчислявате всеки от тях;
3. да обяснявате защо два интензивни показателя не бива да се сравняват пряко, когато групите имат различен състав;
4. да извършвате **пряка стандартизация** в четири стъпки и да формулирате заключението.

1.1 Абсолютни величини

Абсолютните величини са числа, които количествено характеризират обема на една статистическа съвкупност или на нейни части, както и стойността на конкретен статистически признак.

Те имат две задължителни свойства: винаги са **наименовани** — придружени са от мерна единица — и винаги се получават чрез пряко измерване или преброяване.

Систолното артериално налягане, измерено в mmHg, е абсолютна величина: има числова стойност, има мерна единица и характеризира количествено определен признак. Броят на преминалите през едно отделение пациенти за годината също е абсолютна величина.

Важно

Статистическото изследване обикновено **започва** с абсолютни величини, но те **не са достатъчни** за сравнения във времето и пространството. Не можем да кажем, че болница с 400 починали работи по-зле от болница с 40 починали, преди да разберем колко пациенти са преминали през всяка от тях.

1.2 Относителни величини

Относителните величини се получават при деление на две абсолютни величини. Представят се като коефициент, като процент (при умножение по 100), като промил (при умножение по 1 000) или на 100 000 население.

Пример от клиничната практика. Ударният обем е количеството кръв, което постъпва в аортата след едно сърдечно съкращение — абсолютна величина в милилитри. Сърцето на едър състезател изтласква повече милилитри от сърцето на първокласник, но това не означава, че работи по-добре. Затова се използва **фракцията на изтласкване** — отношението на ударния обем към обема на камерата преди съкращението. Тя е относителна величина и именно тя позволява сравнение между двамата.

В медицинската статистика се разграничават два основни вида относителни показатели — екстензивни и интензивни.

1.2.1 Екстензивни показатели

Екстензивните показатели се наричат още *структурни*. Те показват как едно цяло се разпределя между взаимно изключващи се съставни части в определено време и място. Знаменателят е общият брой единици в разглежданото цяло, а числителят е броят в една от неговите категории.

$$\text{екстензивен показател} = \frac{\text{част от явлението}}{\text{цялото явление}} \times 100$$

> **!** Свойство, което ще ползвате постоянно
>
> Сумата от всички екстензивни показатели за едно явление винаги е равна на **1 (100 %)**. Ако при вас не се получава 100 %, някъде има грешка.

Ако разделим болните от хепатит А по възрастови групи и разделим броя във всяка група на общия брой болни, получаваме възрастовата **структура** на заболелите — поредица от екстензивни показатели, чиято сума е 100 %.

1.2.2 Интензивни показатели

Интензивните показатели се наричат още *честотни*. Те показват колко често възниква или се наблюдава дадено събитие в популацията, в която то е възможно. В числителя е броят събития, а в знаменателя — броят лица или човеко-времето под риск. За разлика от екстензивния показател, тук не описваме структурата на едно цяло, а честотата на събитие в неговата среда.

$$\text{интензивен показател} = \frac{\text{брой случаи на явлението}}{\text{обем на средата, в която възниква}} \times 100$$

Четири примера, които трябва да можете да различавате:

- **Леталитет** — брой починали от дадено заболяване спрямо броя **болни** от него. Леталитет 5 % от морбили при деца означава, че от 100 заболели деца 5 са починали.

- **Смъртност** — брой починали спрямо **цялото население** (или населението в риск). Разликата с леталитета е в знаменателя: при леталитета в знаменателя са болните, при смъртността — всички.
- **Заболеваемост** — брой **новозаболели** спрямо популацията в риск.
- **Честота на екзацербации** при астма — брой обостряния спрямо броя проследени пациенти или спрямо натрупаното време на проследяване. Периодът и знаменателят трябва да бъдат посочени изрично.

> Как да ги различите за секунда
>
> Погледнете какво означава **знаменателят**. Ако той е цялото, което разделяме на категории, показателят е **екстензивен**. Ако знаменателят е популацията под риск, в която броим събития, показателят е **интензивен**. Самото обстоятелство, че лицата в числителя се срещат и в знаменателя, не е достатъчно: при леталитета починалите са сред заболелите, но показателят описва честота на изход сред болни, а не структура по категории.

1.3 Защо интензивните показатели не се сравняват пряко

Стойността на един интензивен показател зависи от **структурата на средата**, в която е измерен. Раждаемостта е по-висока там, където преобладава население на възраст 20–30 г. — не защото хората там са по-плодовити, а защото такъв е съставът на населението. Леталитетът в една болница е по-висок, ако през нея преминават предимно онкологични пациенти.

Затова, когато сравняваме интензивни показатели, изчислени в популации с различна структура, общите стойности може да ни подведат. Възрастта или друг структурен признак се смесва с изследваното различие и затруднява сравнението. Методът, с който уеднаквяваме тази структура, се нарича **стандартизация**.

> Определение
>
> **Стандартизацията** е преобразуване на общите (сумарните) коефициенти, с което се отстранява влиянието на възрастовите или други различия в състава на сравняваните групи. Стандартизираните показатели показват какви биха били общите коефициенти, ако сравняваните групи имаха **еднакъв състав**.

В курса се разглежда само **прекият метод** на стандартизация.

1.4 Решен пример

Условие

Изследвани са две групи миньори — **неконсумиращи** и **консумиращи** алкохол. За всяка възрастова група са известни броят преминали профилактичен преглед N и броят заболели C.
Задача. Да се сравнят *стандартизираните* показатели за заболеваемост в двете групи, като за **стандарт** се приеме възрастовата структура на втората група (консумиращите алкохол).

Началните данни са представени в Таблица 1.1.

Таблица 1.1: Начални данни. N_1, C_1 — преминали и заболели сред неконсумиращите алкохол; N_2, C_2 — сред консумиращите

Възраст	N_1	C_1	N_2	C_2
< 24 г.	500	250	4 000	2 400
25–34 г.	500	300	2 000	1 400
35–44 г.	1 000	700	1 000	800
45+ г.	2 000	1 600	1 000	900
Всичко	**4 000**	**2 850**	**8 000**	**5 500**

1.4.1 Стъпка 1. Изчисляване на нестандартизираните показатели

Изчисляваме заболеваемостта поотделно за всяка възрастова група и общо за всяка от двете групи:

$$IR = \frac{C}{N} \times 100$$

За миньорите под 24 г., неконсумиращи алкохол:

$$IR_1 = \frac{250}{500} \times 100 = 50\,\%$$

Общо за неконсумиращите:

$$IR_1^{\text{общо}} = \frac{2\,850}{4\,000} \times 100 = 71{,}25\,\%$$

Общо за консумиращите:

$$IR_2^{\text{общо}} = \frac{5\,500}{8\,000} \times 100 = 68{,}75\,\%$$

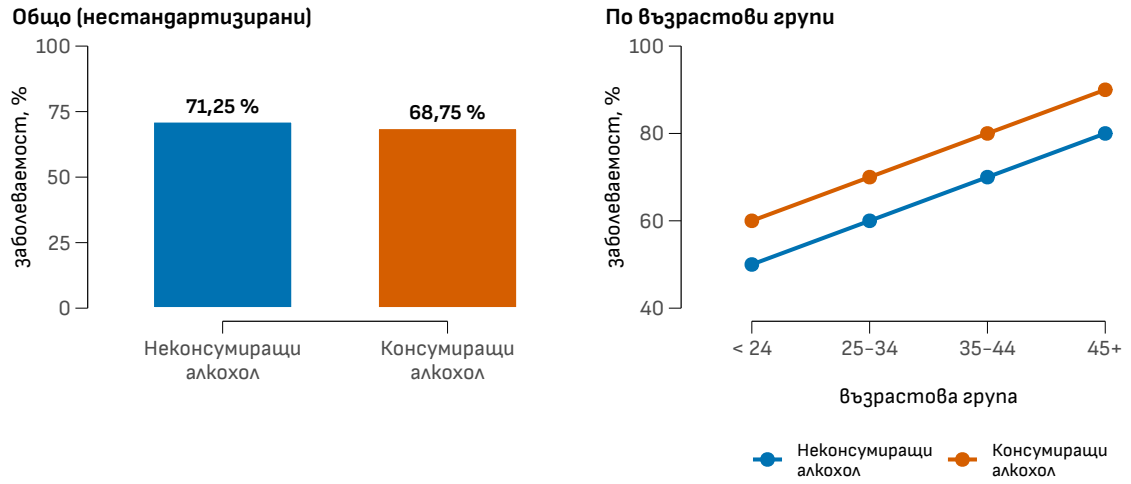
Пълните резултати са в Таблица 1.2.

Таблица 1.2: Стъпка 1 — нестандартизирани показатели за заболеваемост

Възраст	N_1	C_1	IR_1, %	N_2	C_2	IR_2, %
< 24 г.	500	250	50,00	4 000	2 400	60,00
25–34 г.	500	300	60,00	2 000	1 400	70,00
35–44 г.	1 000	700	70,00	1 000	800	80,00
45+ г.	2 000	1 600	80,00	1 000	900	90,00
Всичко	**4 000**	**2 850**	**71,25**	**8 000**	**5 500**	**68,75**

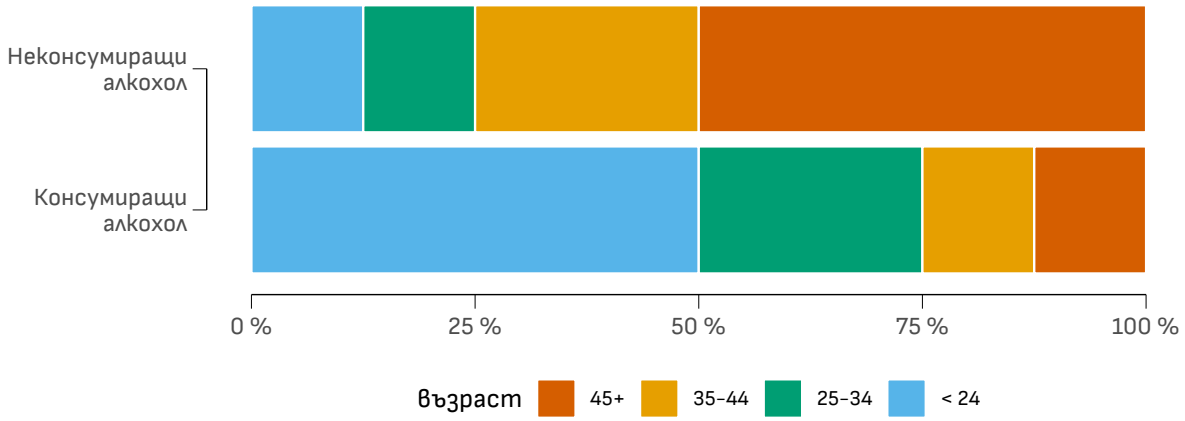
> **!** Важно
>
> Общият **нестандартизиран** показател **не е** сума от показателите по подгрупи. Той се изчислява отново, от общите числа: $2\,850/4\,000$. Ако съберете 50 + 60 + 70 + 80, получавате 260 % — безсмислица.

Резултатът от стъпка 1 изглежда парадоксално и е представен на Фиг. 1.1. Погледнати общо, неконсумиращите алкохол миньори боледуват **по-често** (71,25 % срещу 68,75 %). Погледнати по възрастови групи обаче, консумиращите алкохол боледуват по-често във **всяка една** възрастова група.



Фиг. 1.1: Вляво — общите нестандартизирани показатели: неконсумиращите изглеждат по-болни. Вдясно — същите данни по възрастови групи: консумиращите са по-болни във всяка една група. Противоречието се дължи на различната възрастова структура.

Обяснението е във възрастовата структура (Фиг. 1.2): сред неконсумиращите преобладават **по-възрастните** миньори, а сред консумиращите — **по-младите**. Възрастта е свързана и с фактора, и с резултата — тя е **смесващ фактор**.



Фиг. 1.2: Възрастова структура на двете групи. Сред неконсумиращите алкохол половината миньори са над 45 г., докато сред консумиращите половината са под 24 г. Именно това различие изкривява сравнението на общите показатели.

1.4.2 Стъпка 2. Изчисляване на стандарта

Стандартът е **възрастовата структура**, която приемаме за обща за двете групи. Тук по условие това е структурата на втората група (консумиращите алкохол):

$$S = \frac{N_2^{\text{група}}}{N_2^{\text{общо}}} \times 100$$

За възрастовата група под 24 г.:

$$S_{<24} = \frac{4\,000}{8\,000} \times 100 = 50\ \%$$

Таблица 1.3: Стъпка 2 — изчисляване на стандарта

Възраст	N_2	S, %
< 24 г.	4 000	50,0
25–34 г.	2 000	25,0
35–44 г.	1 000	12,5
45+ г.	1 000	12,5
Всичко	**8 000**	**100,0**

> **Проверка**
>
> Стандартът винаги е **екстензивен показател** — сумата му трябва да е точно 100 %. Ако не е, върнете се и проверете деленето.

Кой стандарт да се избере? В учебните задачи той обикновено е зададен. В реално сравнение се предпочита обща външна популация или сборът на сравняваните групи, за да бъдат

резултатите съпоставими. Избраният стандарт влияе върху конкретните стойности и понякога може да повлияе и върху извода; затова винаги се посочва изрично. В Секция 1.5 ще проверим какво се получава в настоящия пример при друг стандарт.

1.4.3 Стъпка 3. Изчисляване на стандартизираните показатели

За всяка възрастова група умножаваме **нестандартизирания** показател по **квотата по стандарт** за същата възрастова група:

$$AIR = \frac{IR \times S}{100}$$

За неконсумиращите алкохол под 24 г.:

$$AIR_1 = \frac{50 \times 50}{100} = 25\ \%$$

За консумиращите алкохол под 24 г.:

$$AIR_2 = \frac{60 \times 50}{100} = 30\ \%$$

Таблица 1.4: Стъпка 3 — стандартизирани показатели за заболеваемост

Възраст	IR_1, %	IR_2, %	S, %	AIR_1, %	AIR_2, %
< 24 г.	50	60	50,0	25,00	30,00
25–34 г.	60	70	25,0	15,00	17,50
35–44 г.	70	80	12,5	8,75	10,00
45+ г.	80	90	12,5	10,00	11,25
Всичко	**71,25**	**68,75**	**100,0**	**58,75**	**68,75**

! Двете правила за сумиране

- Нестандартизираните показатели **НЕ** се събират.
- Стандартизираните показатели **СЕ** събират — сборът им е общият стандартизиран показател.

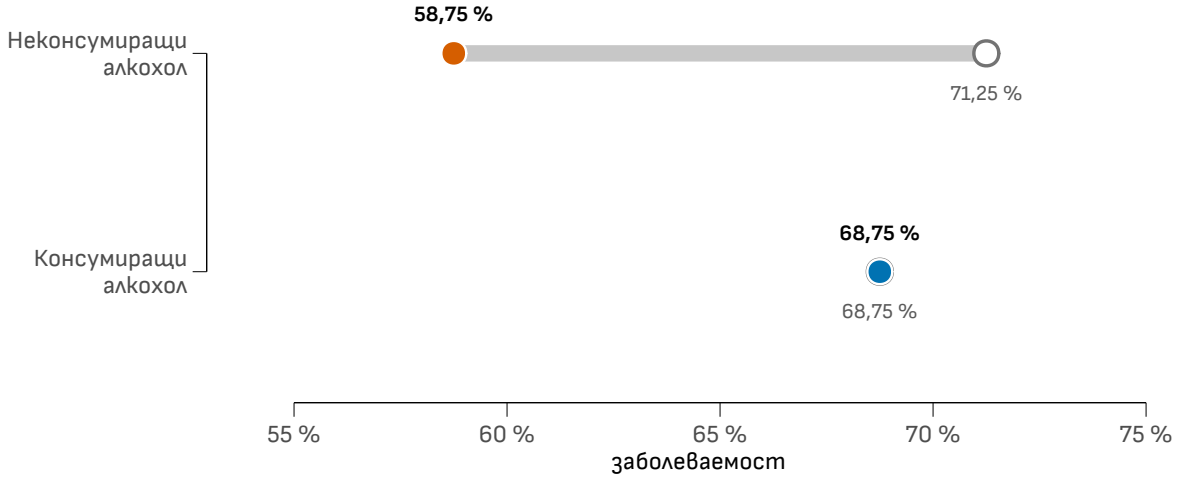
i Полезна проверка

Групата, чиято структура е приета за стандарт, **запазва** стойността на общия си показател преди и след стандартизацията. Тук стандартът е структурата на консумиращите и техният показател остава 68,75 % и в двете колони. Ако това не се получи, търсете аритметична грешка.

1.4.4 Стъпка 4. Заключение

Стандартизираните показатели за заболеваемост в двете групи са съответно **58,75 %** (неконсумиращи) и **68,75 %** (консумиращи). Заболеваемостта сред миньорите, консумиращи алкохол, е с **10 процентни пункта по-висока**.

Изводът е **обратен** на този от нестандартизираните данни. След уеднаквяване на възрастовата структура сравнението показва по-висок стандартизиран показател при консумиращите алкохол. Само по себе си това сравнение не доказва, че алкохолът е причина за различието.



Фиг. 1.3: Преди и след стандартизация. Нестандартизираните показатели подсказват, че неконсумиращите боледуват по-често; след премахване на влиянието на възрастта картината се обръща.

> ⚠ Как се интерпретират стандартизираните показатели
>
> Стандартизираните показатели **не са реални честоти**. Никой не е наблюдавал 58,75 % заболеваемост сред неконсумиращите алкохол — действително наблюдаваната е 71,25 %. Стандартизираният показател отговаря на въпроса: *„каква би била заболеваемостта, ако двете групи имаха еднаква възрастова структура?“* Затова те се използват **само за сравнение помежду си**, а не като самостоятелна оценка на честотата.

1.5 Проверка: същата задача с другия стандарт

За да проверим чувствителността на резултата към избора на стандарт, решаваме отново, този път със структурата на **първата** група (неконсумиращите) като стандарт.

Таблица 1.5: Същата задача, но със структурата на неконсумиращите като стандарт

Възраст	N_1	S, %	IR_1, %	IR_2, %	AIR_1, %	AIR_2, %
< 24 г.	500	12,5	50	60	6,25	7,50
25–34 г.	500	12,5	60	70	7,50	8,75
35–44 г.	1 000	25,0	70	80	17,50	20,00
45+ г.	2 000	50,0	80	90	40,00	45,00
Всичко	**4 000**	**100,0**	**71,25**	**68,75**	**71,25**	**81,25**

Числата са различни — **71,25 %** срещу **81,25 %**, — но разликата отново е **10 процентни пункта** и посоката е същата: консумиращите алкохол боледуват по-често. Забележете и че сега първата група запазва своя показател (71,25 %), защото нейната структура е стандартът.

> Какво да запомните от проверката
>
> В този пример изборът на стандарт променя **конкретните стойности**, но не и **извода**, защото показателят при консумиращите е с 10 процентни пункта по-висок във всяка възрастова група. Това не е универсално правило. Ако разликите по подгрупи са с различна посока, изборът на стандарт може да промени и крайното сравнение. Поради това стандартизиран показател никога не се съобщава без използвания стандарт.

1.6 Обобщение на метода

Пряката стандартизация в четири стъпки:

1. **Нестандартизирани показатели** — за всяка подгрупа поотделно и общо за всяка от сравняваните групи.
2. **Стандарт** — структурата на едната група (или външна популация); сборът на квотите е 100 %.
3. **Стандартизирани показатели по подгрупи** — $AIR = IR \times S/100$.
4. **Сумиране и заключение** — сборът на стандартизираните показатели е общият стандартизиран показател; сравняват се двата общи показателя.

1.7 Чести грешки

Таблица 1.6: Чести грешки при стандартизацията

#	Грешка	Как да я избегнете
1	Сравняване на нестандартизирани показатели при групи с различна структура	Винаги първо погледнете структурата на групите

#	Грешка	Как да я избегнете
2	Сумиране на нестандартизирани показатели, за да се получи общият	Общият нестандартизиран показател се смята наново, от общите числа
3	Събиране вместо умножение в стъпка 3	$AIR = IR \times S/100$ — **умножение**
4	Стандарт, чиито квоти не дават 100 %	Проверявайте сбора преди да продължите
5	Интерпретиране на стандартизирания показател като реална честота	Той е „какво би било, ако…“, а не наблюдение
6	Смесване на леталитет и смъртност	Гледайте знаменателя: болни или население

1.8 Задачи за самостоятелна работа

Задача 1.1

Да се определи възрастовата структура на преминалите болни от хепатит А през болница Х.

Възраст	0–1 г.	1–4 г.	5–9 г.	10–19 г.	Общо
Болни от хепатит А	70	41	102	87	300
Възрастова структура, %					

Какъв вид показател изчислявате и как ще проверите резултата си?

Задача 1.2

Да се изчислят стандартизираните показатели за плодовитост в райони А и Б. За стандарт да се приеме възрастовият състав на жените от район Б.

Възраст	Жени (А)	Живородени (А)	Жени (Б)	Живородени (Б)
15–20 г.	1 000	18	1 200	22
21–30 г.	9 000	225	7 000	175
31–49 г.	8 000	128	10 000	160
Всичко	**18 000**	**371**	**18 200**	**357**

Задача 1.3

Да се изчислят стандартизираните показатели за леталитет в две градски болници. За стандарт да се приеме съставът на болните в болница Б.

Болест	Преминали (А)	Починали (А)	Преминали (Б)	Починали (Б)
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**

Задача 1.4

Нестандартизираните показатели за леталитет в болници А и Б са съответно 10 ‰ и 12 ‰. След стандартизация по структурата на болница А леталитетът в болница А е 10 ‰, а в болница Б — 8 ‰.
Какъв извод следва? Ако за стандарт се използва структурата на болница Б, могат ли новите стандартизирани стойности да се изчислят само от дадената информация? Обосновете отговора си.

Задача 1.5 (за разсъждение)

Общата смъртност в Япония е по-висока от общата смъртност в Бангладеш. Означава ли това, че здравеопазването в Япония е по-лошо? Обяснете отговора си с термините от това упражнение.

Задача 1.6

През една година в болница са лекувани 240 пациенти: 96 в кардиологично, 84 в хирургично и 60 в неврологично отделение. Починали са 18 от всички лекувани пациенти.

1. Изчислете структурата на пациентите по отделения.
2. Изчислете болничния леталитет.
3. Посочете кой от двата резултата е екстензивен и кой — интензивен показател. За всеки резултат обяснете какво означава знаменателят.

1.9 Кратък тест

За всеки въпрос изберете **един** верен отговор. Ключът е в Секция Г.1.

1. Кой показател е екстензивен?

   а) 28 нови случая на 100 000 души за година;
   б) 35 % от хоспитализираните са приети по спешност;
   в) 4 починали на 100 заболели;
   г) 12 раждания на 1 000 жени във фертилна възраст.

2. Основната разлика между смъртност и леталитет е:

   а) мерната единица на числителя;
   б) периодът на наблюдение;
   в) знаменателят на показателя;
   г) възрастта на изследваните лица.

3. Сумата на квотите в избраната стандартна структура трябва да бъде:

   а) 1 %;
   б) 10 %;
   в) 100 %;
   г) равна на броя подгрупи.

4. Общият стандартизиран показател се получава като:

   а) сбор на нестандартизираните показатели по подгрупи;
   б) сбор на стандартизираните показатели по подгрупи;
   в) средна аритметична на общите наблюдавани показатели;
   г) произведение на стандартизираните показатели.

5. Кое твърдение за стандартизирания показател е вярно?

   а) Той е непосредствено наблюдавана честота;
   б) Може да се сравнява независимо от използвания стандарт;
   в) Показва какъв би бил общият показател при зададена обща структура;
   г) Доказва причинна връзка между фактора и резултата.

2 Дескриптивна статистика

Цел на упражнението

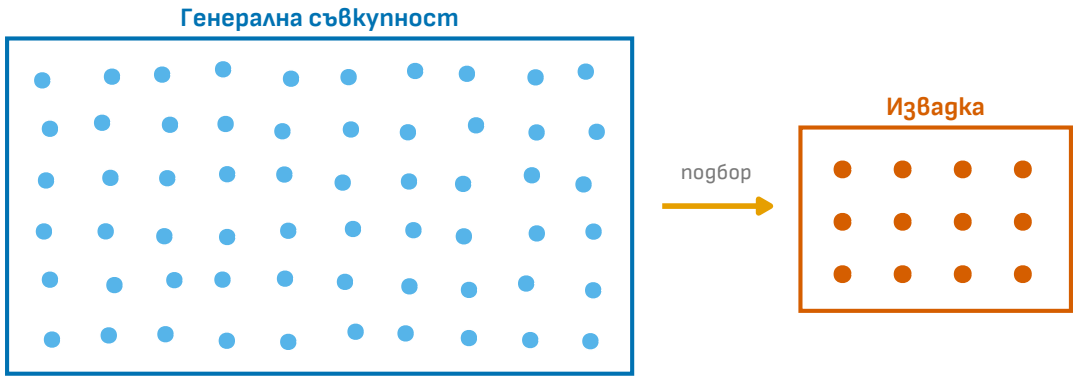
След това упражнение трябва да умеете:

1. да определяте вида на статистическия признак и скалата, на която се измерва;
2. да групирате данни в **степенен** и в **интервален** вариационен ред;
3. да изчислявате средна аритметична, медиана и мода при негрупирани и при групирани данни;
4. да изчислявате размах, стандартно отклонение и интерквартилен размах;
5. да избирате правилната двойка показатели: $\bar{X} \pm SD$ или Me (IQR).

2.1 Извадка и генерална съвкупност

Едно статистическо изследване изисква време, пари и хора. Почти никога не е възможно да се изследват всички индивиди, които ни интересуват. Немислимо е да се прегледат всички пациенти със сърдечна недостатъчност, за да се установи дали ново лекарство намалява пулса им.

Затова се изследва по-малка част — **извадка**, — а чрез статистически методи се правят изводи за всички лица, към които е насочен въпросът — **генералната съвкупност** (Фиг. 2.1).



Фиг. 2.1: От генералната съвкупност се подбира извадка. Стрелката представя процедурата за подбор, а не преместване на едни и същи лица.

За да бъдат тези изводи надеждни, извадката трябва да е получена чрез подходяща процедура за подбор, да обхваща правилно определена целева популация и да няма съществено систематично отпадане. Случайният подбор намалява риска от систематична грешка, но не гарантира, че всяка извадка ще възпроизведе точно структурата на популацията по пол, възраст или друг признак.

⚠ Двете половини на статистиката

Дескриптивната (описателната) статистика описва какво е установено **в извадката**.
Инферентната статистика пренася този резултат върху **генералната съвкупност**. С нея се занимаваме в Секция 3.

2.2 Статистически признаци и скали за измерване

Статистическите данни са числови или буквени стойности, записани след измерване или преброяване на определена характеристика. Тъй като тези характеристики приемат различни стойности при различните лица, те се наричат **променливи**. Ако характеристиката е една и съща за всички — тя е **константа**.

⚠ Определение

Статистическият признак е отличителна черта, белег на една или повече единици на наблюдение.

Признаците се делят на две основни групи.

Количествените признаци се изразяват с числа, за които аритметичните операции имат смисъл. Те биват:

- *прекъснати (дискретни)* — приемат отделни, изолирани стойности: брой живородени деца, брой хоспитализации за година;
- *непрекъснати (континуални)* — могат да се измерят с произволна точност след десетичната запетая: тегло, температура, кръвно налягане.

Категориалните (качествените) признаци разпределят наблюденията в категории. За тях се използват етикети, например кръвна група, вид заболяване или наличие на алергия. Ако категориите са кодирани с числа, кодовете остават етикети и не се превръщат автоматично в количествени стойности.

Скалата, на която се измерва един признак, определя кои статистически методи са допустими (Таблица 2.1).

Таблица 2.1: Видове признаци и скали на измерване

Скала	Характеристика	Пример
Номинална	категории, „наименования, етикети“; няма ред	диагноза, кръвна група
Дихотомна	*вид номинална скала само с две възможни стойности*	*пол; жив/починал*

Скала	Характеристика	Пример
Ординална	категории с логически ред (нарастващ или намаляващ); разликите не са измерими	степен на сърдечна недостатъчност (лека, умерена, тежка); стадий на онкологично заболяване
Рангова	*вид ординална скала — единиците получават поредни места от 1 до n*	*класиране по успех*
Интервална	числа; стойността „0“ **не** означава липса на признака; разлики — да, отношения — не	температура по Целзий
Пропорционална	числа; стойността „0“ означава липса на признака; допуска и отношения	тегло, ръст, брой дни

Номиналната и ординалната скала, включително техните разновидности, са **неметрични**. Интервалната и пропорционалната са **метрични**. Определенията „слаби“ и „силни“ описват какви операции допуска скалата, а не дали измерването е извършено добре или зле.

> Защо това има значение
>
> Ако кажем за човек, че е „висок“, това не е същото като „ръст 185 см“. В първия случай сме измерили на слаба скала, във втория — на силна. Колкото по-силна е скалата, толкова повече информация носи признакът и толкова повече методи са допустими.

2.3 Показатели за централна тенденция

Дескриптивният анализ описва извадката чрез две групи обобщаващи числови характеристики: **централна тенденция** и **разсейване**.

Централната тенденция е стойност, която представя всички наблюдения „като едно цяло“ — типичният очакван представител на извадката. Разсейването показва колко са разпръснати наблюденията около нея.

За променливи, измерени на **интервална** или **пропорционална** скала, могат да се използват средната аритметична и медианата, като изборът зависи и от формата на разпределението. При **номинални** променливи се съобщават честоти, относителни дялове и мода. При **ординални** променливи могат да се използват и медиана, и квартили, когато подредбата на категориите го позволява.

2.3.1 Средна аритметична при негрупирани данни

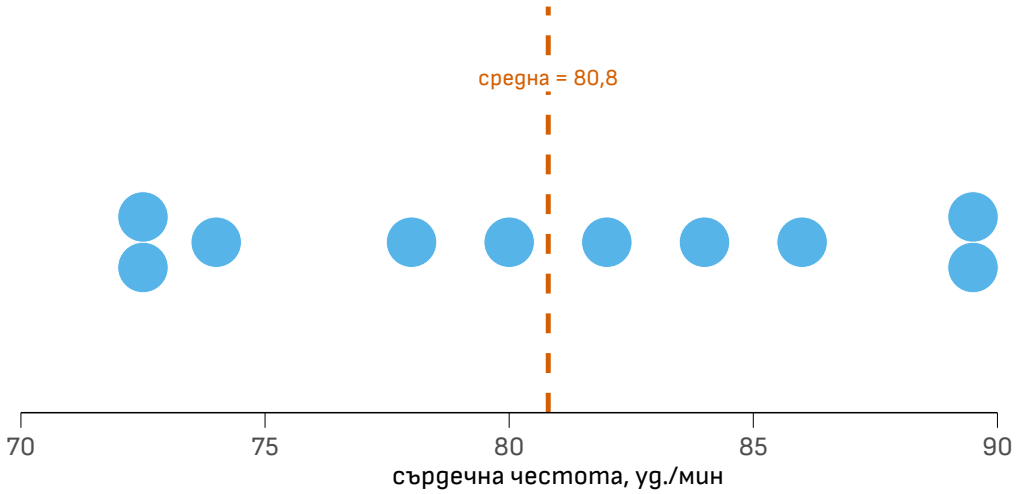
$$\bar{X} = \frac{\sum X}{n}$$

където $\bar{X}$ е средната аритметична, $\sum X$ — сборът на всички наблюдавани стойности, а n — броят на наблюденията.

Пример

Сърдечната честота на 10 студенти от II курс е:
90, 80, 89, 72, 73, 74, 78, 86, 84, 82 уд./мин.

$$\bar{X} = \frac{90 + 80 + 89 + 72 + 73 + 74 + 78 + 86 + 84 + 82}{10} = \frac{808}{10} = 80{,}8 \text{ уд./мин}$$



Фиг. 2.2: Разпределение на пулса при 10 студенти. Всяка точка е едно наблюдение; еднаквите или близките стойности се подреждат една над друга. Прекъснатата линия е средната аритметична.

2.3.2 Групиране на данните

При по-голям обем данни е удобно те да се представят групирани. Множеството от различните стойности (или интервали) на признака заедно с техните честоти се нарича **честотно разпределение**.

Степенен вариационен ред. Записва се всяка отделна стойност на признака и срещу нея — броят наблюдения с тази стойност. При него **не се губи** статистическа точност: знаем стойността на всяко наблюдение.

Интервален вариационен ред. Стойностите се групират в интервали. При него се губи точност — знаем само в кой интервал попада наблюдението, но не и точната му стойност.

Две правила при съставяне на интервален ред

1. Интервалите са с **еднаква ширина**.
2. Всяко наблюдение попада в **точно един** интервал. За целочислени данни може да се използва записът 1–5, 6–10, 11–15. За непрекъснати данни границите се задават еднозначно, например $[0; 5)$, $[5; 10)$ и $[10; 15]$.

2.3.3 Средна аритметична при групирани данни

$$\bar{X} = \frac{\sum (X_u \cdot f)}{\sum f}$$

където:

- X_u е **средата на интервала**: $X_u = \frac{\text{долна граница} + \text{горна граница}}{2}$;
- f е честотата — броят наблюдения в интервала;
- $\sum f = n$ е обемът на извадката.

Основен пример за упражнението

Проследени са $n = 30$ пациенти, лекувани с експериментална терапия. Преживяемостта в месеци е:
3, 11, 6, 9, 10, 10, 12, 15, 1, 4, 8, 12, 9, 9, 6, 5, 9, 10, 7, 11, 10, 12, 12, 6, 4, 15, 11, 9, 4, 5.
Групирани в интервален ред, данните изглеждат така:

Преживяемост X, месеци	Честота f
$1 - 5$	7
$6 - 10$	14
$11 - 15$	9
Общо	**30**

Стъпка 1 — намираме средата на всеки интервал и произведението $X_u \cdot f$:

Таблица 2.2: Изчисляване на средната аритметична при интервален ред

Преживяемост X	f	X_u	$X_u \cdot f$
$1 - 5$	7	3	21
$6 - 10$	14	8	112
$11 - 15$	9	13	117
Общо	**30**		**250**

Стъпка 2 — прилагаме формулата:

$$\bar{X} = \frac{250}{30} = 8{,}33 \text{ месеца}$$

Групирането „струва“ малко точност

Ако изчислим средната от изходните 30 стойности, получаваме 8,5 месеца. Групираната оценка дава 8,33 месеца. Разликата от 0,17 месеца е цената на групирането — всички наблюдения в един интервал са заменени със средата му.

2.3.4 Медиана и мода

Медианата Me е стойността, която разделя възходящо подредения статистически ред на две равни части: 50 % от наблюденията са под нея и 50 % — над нея.

- При **нечетен** брой наблюдения медианата е стойността на средното наблюдение. За реда 6,0; 6,5; **7,0**; 7,5; 8,0 медианата е 7,0.
- При **четен** брой медианата е средната аритметична на двете средни стойности. За реда 6,0; 6,5; **7,0; 7,5**; 8,0; 8,5 медианата е $(7{,}0 + 7{,}5)/2 = 7{,}25$.

Модата M_0 е най-често срещаната стойност. Намира се чрез броене. За реда 120; 120; 130; 135; 120; 140; 140 модата е 120. Възможно е да няма мода или да има повече от една. В медицинската статистика модата се използва рядко.

2.3.5 Относителен дял

За качествени признаци централната тенденция се описва с **относителния дял**:

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

където m е броят наблюдения с интересуващата ни характеристика, а n — броят на всички наблюдения.

За група от 10 студенти, 4 от които са момичета:

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

2.4 Показатели за разсейване

2.4.1 Размах

Размахът R е разликата между максималната и минималната стойност:

$$R = X_{max} - X_{min}$$

За пулса на 10-те студенти: $R = 90 - 72 = 18$ уд./мин.

Размахът се използва рядко, защото зависи само от две стойности, а те често са екстремни или дори погрешно измерени.

2.4.2 Стандартно отклонение при негрупирани данни

Стандартното отклонение S (или SD) е мярка за средното ниво на разсейване спрямо средната аритметична:

$$S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n - 1}}$$

където:

- X_i е стойността на i-тото наблюдение;
- $X_i - \bar{X}$ е **отклонението** на наблюдението от средната;
- $(X_i - \bar{X})^2$ е квадратът на отклонението — квадратирането премахва знаците, така че положителните и отрицателните отклонения да не се занулят взаимно;
- $n - 1$ е **корекцията на Бесел**: използва се, когато работим с извадка, а не с цялата популация;
- коренуването връща резултата в първоначалните мерни единици.

Пример — пулсът на 10-те студенти

Средната аритметична е 80,8 уд./мин, но нито един студент няма точно тази стойност. Всеки се отклонява с няколко удара под или над нея.

Пулс	Отклонение $X_i - \bar{X}$	$(X_i - \bar{X})^2$
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
Общо	**0,0**	**383,60**

$$S = \sqrt{\frac{383{,}60}{10 - 1}} = \sqrt{42{,}62} = 6{,}53 \text{ уд./мин}$$

Резултатът се записва като $\bar{X} \pm S = 80{,}8 \pm 6{,}53$ уд./мин.

Защо квадратираме

Сборът на отклоненията винаги е нула — вижте колоната в примера. Ако не квадратирахме, „средното отклонение" щеше да е нула за всяка извадка и показателят щеше

да е безполезен.

2.4.3 Стандартно отклонение при интервален ред

$$S = \sqrt{\frac{\sum (X_u - \bar{X})^2 \cdot f}{n - 1}}$$

Продължаваме примера с 30-те пациенти ($\bar{X} = 8{,}33$ месеца).

Таблица 2.3: Изчисляване на стандартното отклонение при интервален ред

Преживяемост	f	X_u	$X_u - \bar{X}$	$(X_u - \bar{X})^2$	$(X_u - \bar{X})^2 \cdot f$
$1 - 5$	7	3	−5,33	28,44	199,11
$6 - 10$	14	8	−0,33	0,11	1,56
$11 - 15$	9	13	4,67	21,78	196,00
Общо	**30**				**396,67**

$$S = \sqrt{\frac{396{,}67}{30 - 1}} = \sqrt{13{,}68} = 3{,}70 \text{ месеца}$$

Преживяемостта в извадката се описва като $8{,}33 \pm 3{,}70$ месеца.

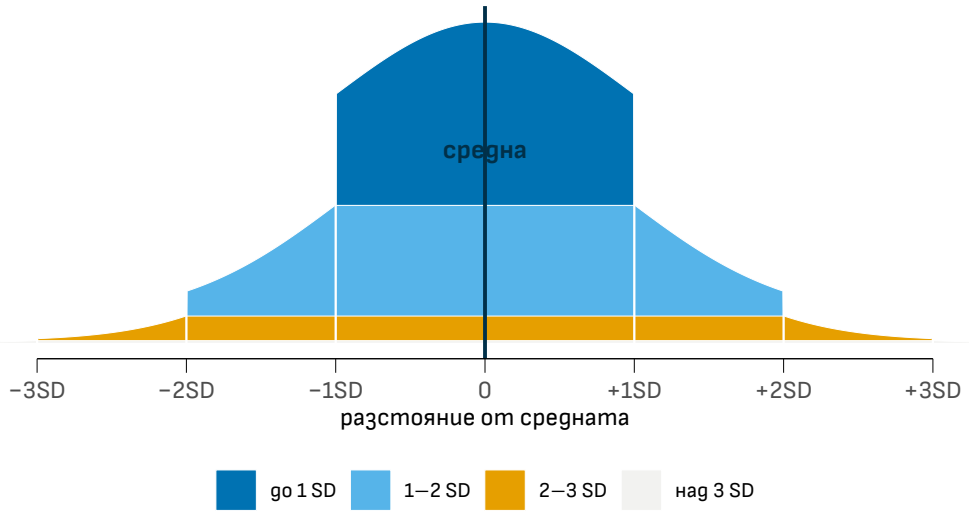
2.4.4 Правилото на трите сигми

Ако признакът е **нормално разпределен** — симетрично, камбановидно разпределение, при което средната, медианата и модата съвпадат (Фиг. 2.3) — стандартното отклонение описва какъв дял от наблюденията се намира на определено разстояние от средната:

- $\bar{X} \pm 1 \cdot S$ обхваща **68 %** от наблюденията;
- $\bar{X} \pm 2 \cdot S$ обхваща **95 %** от наблюденията;
- $\bar{X} \pm 3 \cdot S$ обхваща **99,7 %** от наблюденията.

Пример

Измерено е теглото на извадка от 1 000 души: $\bar{X} \pm S = 80 \pm 10$ кг. Приемаме, че теглото е нормално разпределено.
Тогава 95 % от участниците (950 души) имат тегло в интервала $80 \pm 2 \cdot 10$, тоест **между 60 и 100 кг**.
Останалите 5 % (50 души) се разпределят симетрично в двата края: около 25 души (2,5 %) тежат над 100 кг и около 25 души (2,5 %) — под 60 кг.



Фиг. 2.3: Нормално разпределение и правило на трите сигми. Цветните области показват интервалите от 0 до 1, от 1 до 2 и от 2 до 3 стандартни отклонения от средната; разпределението е симетрично.

> ⚠ Внимание — често се бърка
>
> Правилото на трите сигми описва **как са разпръснати отделните наблюдения**. То не бива да се смесва с доверителния интервал от Секция 3, който описва **колко точно е оценен средният показател**.

2.4.5 Интерквартилен размах

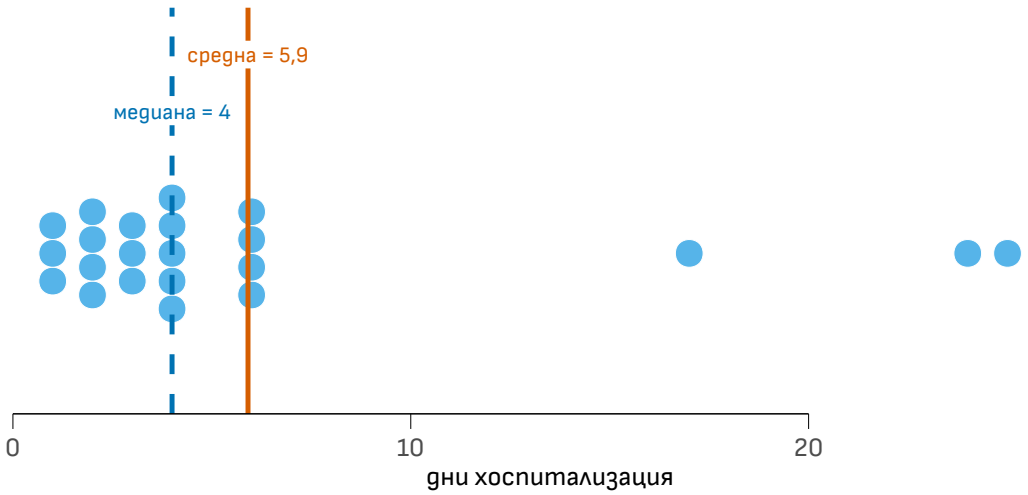
Понякога средната аритметична не описва добре централната тенденция. Проследени са 22 пациенти и е записана продължителността на хоспитализацията им в дни:

24, 4, 2, 3, 4, 1, 6, 17, 4, 3, 2, 6, 25, 1, 6, 4, 2, 4, 3, 6, 2, 1.

Средната аритметична е 5,91 дни. Това число обаче е подвеждащо: само 3 от 22 пациенти (13,6 %) са лежали повече от 6 дни. Техните стойности са **екстремни** и издърпват средната нагоре (Фиг. 2.4).

> ℹ Вид на разпределението и избор на показател
>
> - При **дясно изтеглено** разпределение (дълга опашка надясно) медианата е **по-малка** от средната аритметична.
> - При **ляво изтеглено** разпределение медианата е **по-голяма** от средната.
> - При **нормално** разпределение двете съвпадат или почти съвпадат.
>
> Когато разпределението е изтеглено, се съобщават **медиана и интерквартилен размах**, а не средна и стандартно отклонение.

Интерквартилният размах IQR измерва разсейването спрямо медианата:



Фиг. 2.4: Дясно изтеглено разпределение. Средната аритметична (непрекъсната линия, 5,9 дни) е изместена към екстремните стойности. Медианата (прекъсната линия, 4 дни) описва централната тенденция по-добре.

$$IQR = Q_3 - Q_1$$

където Q_1 и Q_3 са първият и третият квартил — стойностите, които заедно с медианата Q_2 разделят подредения ред на четири равни части.

Подреждаме данните възходящо:

1, 1, 1, 2, 2, 2, **2**, 3, 3, 3, 4, **4**, 4, 4, 4, 6, **6**, 6, 6, 17, 24, 25.

При 22 наблюдения:

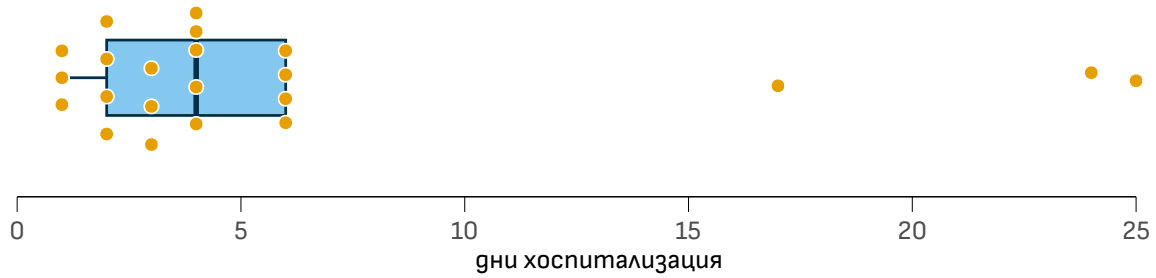
- $Q_1 = 2$ дни — под тази стойност е първата четвърт от пациентите;
- $Q_2 = Me = 4$ дни — медианата;
- $Q_3 = 6$ дни — под тази стойност са три четвърти от пациентите.

$$IQR = Q_3 - Q_1 = 6 - 2 = 4 \text{ дни}$$

Резултатът се записва като $Me\ (IQR) = 4\ (4)$ дни или, по-подробно, $Me = 4$ дни; Q_1–Q_3: 2 – 6 дни.

Как се чете боксплот

- Кутията обхваща средните 50 % от наблюденията — широчината ѝ е IQR.
- Линията в кутията е **медианата**, а не средната аритметична.
- „Мустаците" достигат до най-далечното наблюдение в рамките на $1{,}5 \cdot IQR$ от кутията.
- Точките отвъд мустаците са **екстремни стойности (аутлайъри)**.
- По-дълъг мустак и повече отдалечени точки от едната страна **подсказват** изтегляне към същата страна.



Фиг. 2.5: Боксплот на продължителността на хоспитализацията. Лявата страна на кутията е Q1, дясната — Q3, вертикалната линия вътре е медианата. Точките извън мустаците са екстремни стойности.

Екстремни са всички стойности, по-високи от $Q_3 + 1{,}5 \cdot IQR$ или по-ниски от $Q_1 - 1{,}5 \cdot IQR$. В примера границата е $6 + 1{,}5 \cdot 4 = 12$ дни, тоест стойностите 17, 24 и 25 дни са екстремни.

2.5 Кой показател кога

Таблица 2.4: Избор на показатели според вида на признака

Вид на признака / разпределението	Централна тенденция	Разсейване
Количествен, нормално разпределен	средна аритметична $\bar{X}$	стандартно отклонение S
Количествен, изтеглено разпределение	медиана Me	интерквартилен размах IQR
Качествен (номинален, ординален)	относителен дял $\hat{p}$	—

2.6 Чести грешки

Таблица 2.5: Чести грешки при дескриптивния анализ

#	Грешка	Как да я избегнете
1	Средна аритметична при силно изтеглено разпределение	Погледнете разпределението или боксплота преди да изберете показател
2	Деление на n вместо на $n - 1$ при стандартното отклонение	В извадка **винаги** $n - 1$
3	Забравяне на умножението по f при интервален ред	Всеки интервал „тежи“ според броя наблюдения в него
4	Препокриващи се интервали (1–5, 5–10)	Границите не се повтарят
5	Смесване на $\bar{X} \pm 2S$ с доверителен интервал	Първото описва наблюденията, второто — точността на оценката

#	Грешка	Как да я избегнете
6	Съобщаване на число без мерна единица	„8,33 месеца“, а не „8,33“

2.7 Задачи за самостоятелна работа

Задача 2.1

Измерен е пулсът на 45 студенти. Данните са групирани в интервален ред:

Пулс	48–52	53–57	58–62	63–67	68–72	73–77	78–82
Брой студенти	4	5	8	11	6	6	5

Изчислете средната аритметична и стандартното отклонение.

Задача 2.2

Може ли стандартното отклонение да бъде равно на нула? А отрицателно? Обяснете и в двата случая.

Задача 2.3

Възможно ли е при дадена извадка **всички** индивидуални отклонения $X_i - \bar{X}$ да бъдат положителни? А всички да бъдат нула? Обяснете.

Задача 2.4

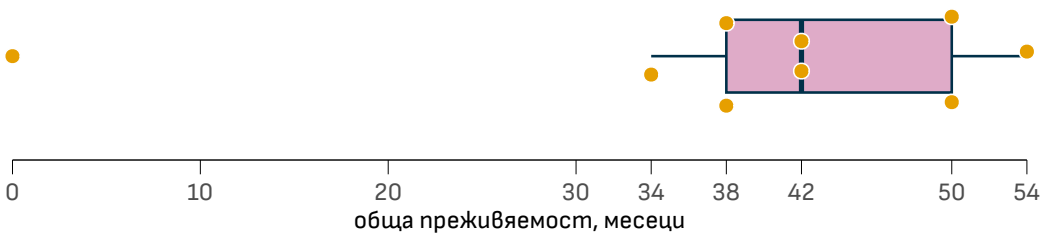
Извадка от 2 000 новородени е включена в проспективно изследване. Средното тегло при раждане е 3 500 г със стандартно отклонение 400 г. Теглото е нормално разпределена величина.
Колко от новородените имат тегло, по-голямо от 4 300 г или по-малко от 3 100 г?

Задача 2.5

Определете вида на признака и скалата, на която се измерва:
а) стадий на бъбречно заболяване (I–V); б) брой цигари на ден; в) кръвна група; г) телесна температура в °C; д) резултат от тест за тревожност (0–100 точки); е) наличие на алергия (да/не).

Задача 2.6

На боксплота по-долу е представена общата преживяемост в месеци.



Фиг. 2.6: Боксплот към задача 2.6.

Каква е стойността на интерквартилния размах? Какъв е видът на разпределението? Има ли екстремни стойности и какви са те?

Задача 2.7

Концентрацията на C-реактивен протеин при девет пациенти е 2, 3, 3, 4, 4, 5, 6, 7 и 18 mg/L.
Изчислете средната аритметична, медианата, Q_1, Q_3 и IQR. Коя двойка показатели описва по-подходящо тези данни и защо?

Задача 2.8

Две малки извадки имат следните стойности:

- група А: 98, 99, 100, 101, 102;
- група Б: 80, 90, 100, 110, 120.

Сравнете средните аритметични и стандартните отклонения. Коя група е по-хомогенна и кой показател подкрепя извода?

2.8 Кратък тест

За всеки въпрос изберете **един** верен отговор. Ключът е в Секция Г.2.

1. Променливата „кръвна група“ се измерва на:

   а) номинална;
   б) ординална;
   в) интервална;
   г) пропорционална скала.

2. Кой показател се влияе най-слабо от единична много висока стойност?

   а) средна аритметична;
   б) размах;
   в) медиана;
   г) стандартно отклонение.

3. При силно дясно изтеглено разпределение най-подходящо е да се съобщят:

   а) средна и стандартно отклонение;
   б) медиана и интерквартилен размах;
   в) мода и размах;
   г) относителен дял и стандартно отклонение.

4. Стандартното отклонение е равно на нула, когато:

   а) средната е нула;
   б) извадката съдържа отрицателни стойности;
   в) всички наблюдения имат една и съща стойност;
   г) обемът на извадката е четно число.

5. При нормално разпределение приблизително 95 % от наблюденията са в:

   а) $\bar{X} \pm 1S$;
   б) $\bar{X} \pm 2S$;
   в) $\bar{X} \pm 3S$;
   г) $\bar{X} \pm 4S$.

6. Линията вътре в кутията на стандартния боксплот показва:

   а) средната;
   б) модата;
   в) медианата;
   г) размаха.

3 Инферентна статистика. Доверителен интервал

Цел на упражнението

След това упражнение трябва да умеете:

1. да различавате **статистика** от **параметър**;
2. да обяснявате защо различни случайни извадки дават различни оценки;
3. да изчислявате стандартна грешка на средната аритметична и на относителния дял;
4. да строите 95 % доверителен интервал за средна и за относителен дял;
5. да формулирате правилно интерпретацията на доверителния интервал;
6. да преценявате как обемът на извадката, разсейването и нивото на доверителност променят ширината на интервала.

Какво трябва да знаете отпреди

От Секция 2: средна аритметична при групирани данни, стандартно отклонение и относителен дял.

3.1 Статистика и параметър

Показателят, изчислен **от извадката**, се нарича **статистика**. Съответният показател за **генералната съвкупност** се нарича **параметър** и обикновено е неизвестен. Извадковата статистика служи като оценка на популационния параметър.

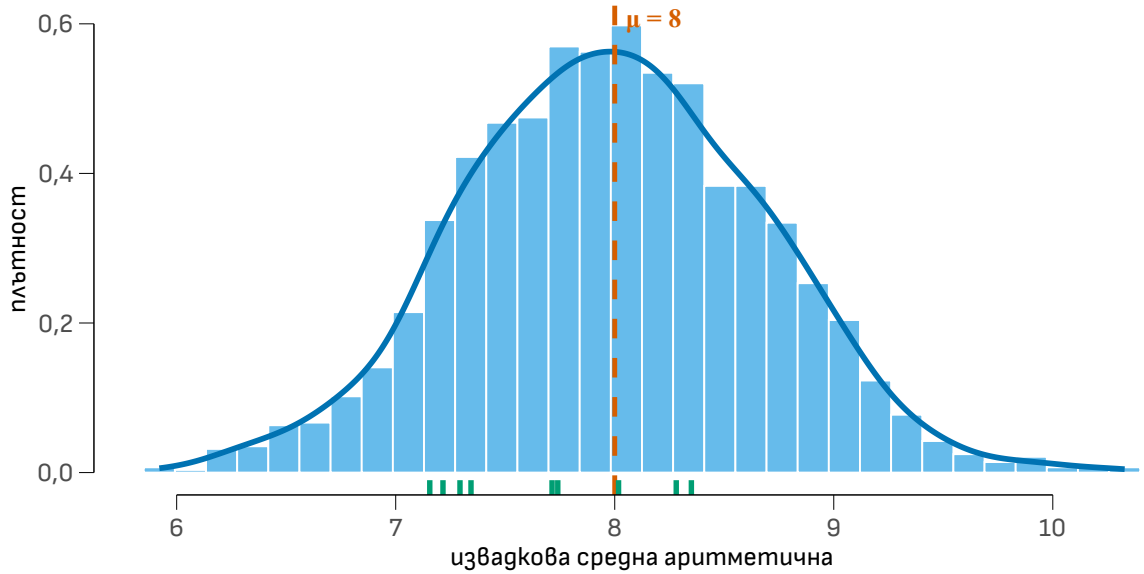
Таблица 3.1: Статистики и параметри

В извадката (статистика)	В генералната съвкупност (параметър)
средна аритметична $\bar{X}$	средна аритметична μ
относителен дял $\hat{p}$	относителен дял P
стандартно отклонение S	стандартно отклонение σ

Средният пулс на 10 случайно подбрани студенти е статистика. Средният пулс на всички студенти от II курс е параметър. Ако вземем нова случайна извадка, вероятно ще получим малко по-различна средна. Това естествено различие между извадките се нарича **случайна извадкова грешка**. То не означава непременно, че измерването или изчислението е погрешно.

3.2 Извадково разпределение и стандартна грешка

Представете си, че многократно избираме случайни извадки с един и същ обем от една генерална съвкупност. За всяка извадка изчисляваме средна аритметична. Получените извадкови средни образуват **извадково разпределение**. Центърът му е около популационната средна μ, а разсейването му показва колко се изменя $\bar{X}$ от извадка на извадка (Фиг. 3.1).



Фиг. 3.1: Извадково разпределение на средната аритметична. Хистограмата показва средните от 2 000 случайни извадки по 30 наблюдения. Прекъснатата линия е популационната средна, а цветните чертички са средните на девет отделни извадки.

Ако изследваният признак е нормално разпределен, извадковото разпределение на средната също е нормално. При достатъчно големи независими извадки то е приблизително нормално и когато разпределението на отделните наблюдения не е нормално. Това е практическото следствие на **централната гранична теорема**. С увеличаването на n извадковите средни се съсредоточават все по-близо до μ.

Стандартната грешка на средната (в някои учебни материали: *средна грешка*) измерва разсейването на извадковите средни и следователно прецизността на $\bar{X}$ като оценка на μ:

$$SE_{\bar{X}} = S_{\bar{X}} = \frac{S}{\sqrt{n}},$$

където S е извадковото стандартно отклонение, а n е обемът на извадката. Малката стандартна грешка не изключва систематична грешка: една пристрастно подбрана извадка може да даде много прецизна, но невярна оценка.

> **!** От какво зависи стандартната грешка
>
> - При по-голямо **стандартно отклонение** наблюденията са по-разпръснати и стандартната грешка е по-голяма.

- При по-голям **обем на извадката** стандартната грешка е по-малка. Тя намалява с $\sqrt{n}$: за да я намалим наполовина, трябва да увеличим извадката четири пъти.

⚠ Не бъркайте стандартното отклонение и стандартната грешка

Стандартното отклонение S описва разсейването на **отделните наблюдения** около средната. **Стандартната грешка** $SE_{\bar{X}}$ описва разсейването на **извадковата средна** от извадка на извадка. При $n > 1$ стандартната грешка е по-малка от стандартното отклонение.

3.3 Доверителен интервал за средна аритметична

Точковата оценка $\bar{X}$ е едно число, но не показва колко е прецизна. **Доверителният интервал** добавя граници около нея и така представя неопределеността на оценката. Общият му вид е:

$$\text{оценка} \pm \text{критична стойност} \times \text{стандартна грешка}.$$

Когато популационното стандартно отклонение σ е известно, се използва критична стойност Z:

$$CI = \bar{X} \pm Z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

На практика σ почти винаги е неизвестно и се заменя с S. Тогава за средна аритметична правилният множител е критична стойност от t-разпределението с $n - 1$ степени на свобода:

$$CI = \bar{X} \pm t_{1-\alpha/2,\, n-1} \cdot \frac{S}{\sqrt{n}}.$$

Този интервал предполага независимо и случайно подбрани наблюдения. При малка извадка е необходимо и разпределението в популацията да е приблизително нормално, без силна асиметрия и крайни стойности. При по-големи извадки методът е по-устойчив на умерени отклонения от нормалността.

При големи извадки t и Z са почти еднакви и често се използва приближението $Z = 1{,}96$. При малки и умерени извадки t е малко по-голямо и дава по-широк интервал. В Таблица 3.2 са дадени стойностите на Z, които се използват и като голямоизвадково приближение.

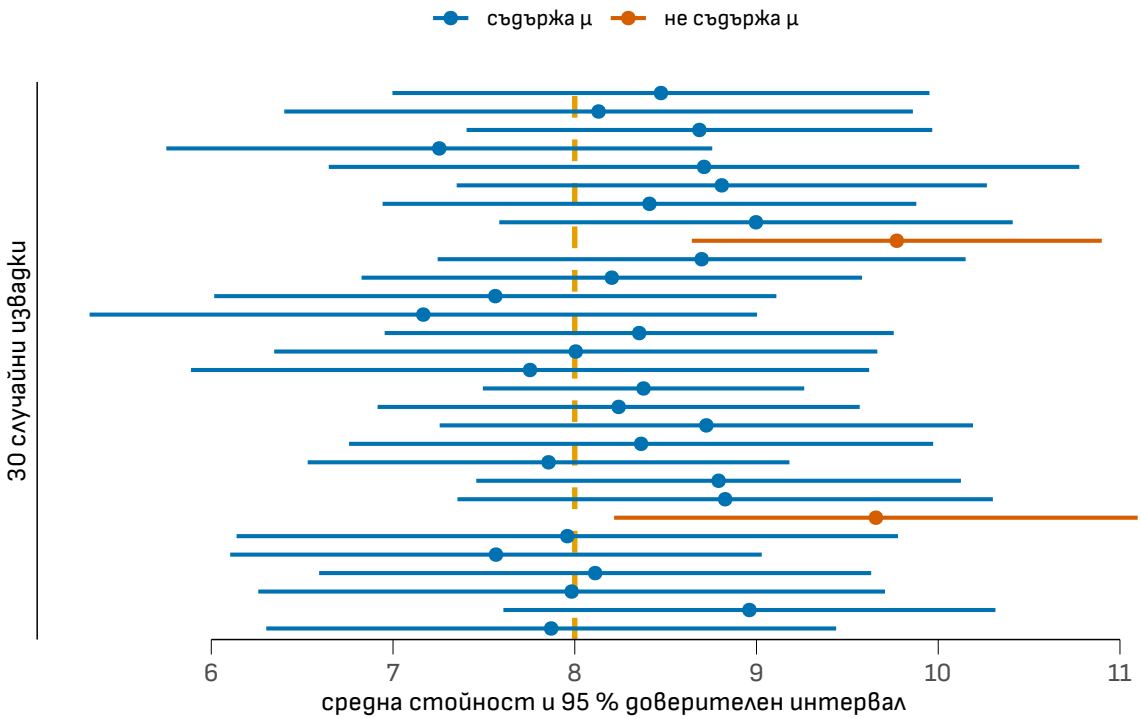
Таблица 3.2: Критични стойности на стандартното нормално разпределение

Ниво на доверителност	α	$Z_{1-\alpha/2}$
90 %	0,10	1,645
95 %	**0,05**	**1,960**

Ниво на доверителност	α	$Z_{1-\alpha/2}$
99 %	0,01	2,576

3.3.1 Какво означава „95 % доверителен“

Преди извадката да бъде избрана, границите на интервала са случайни, защото зависят от случайно подбраните наблюдения. При многократно повторение на процедурата около 95 % от построените по този начин интервали ще съдържат истинския параметър. В отделна серия делът може да не бъде точно 95 % (Фиг. 3.2).



Фиг. 3.2: Тридесет 95 % доверителни интервала за една и съща популационна средна. Сините интервали съдържат истинската стойност μ = 8, а червените не я съдържат. Нивото 95 % описва надеждността на метода в дълга серия, а не обещава точно определен брой успешни интервали във всяка серия.

След като са изчислени конкретните граници, параметърът е една фиксирана, макар и неизвестна стойност. Затова при тази интерпретация не казваме: „има 95 % вероятност параметърът да е в този конкретен интервал“. Казваме: „изчисленият 95 % доверителен интервал е от … до …“ и посочваме, че използваният метод има 95 % покритие при многократно повторение и изпълнени предпоставки.

> **!** Защо не използваме 100 % доверителен интервал
>
> Нетривиален краен интервал с 100 % покритие по правило не може да се построи. Интервалът, който включва всички възможни стойности, би бил сигурен, но практически безполезен. Статистиката не премахва неопределеността, а я измерва.

3.3.2 Решен пример: доверителен интервал за средна

Условие

Проследени са $n = 30$ пациенти, лекувани с експериментална терапия. Преживяемостта им е представена като интервален вариационен ред:

Преживяемост X, месеци	Честота f
1–5	7
6–10	14
11–15	9
Общо	**30**

Задача. Да се изчисли 95 % доверителен интервал за популационната средна преживяемост μ.

Стъпка 1. Извадкова средна аритметична. Средите на интервалите са 3, 8 и 13 месеца.

$$\bar{X} = \frac{3 \cdot 7 + 8 \cdot 14 + 13 \cdot 9}{30} = \frac{250}{30} = 8{,}33 \text{ месеца.}$$

Стъпка 2. Извадково стандартно отклонение. При групирани данни използваме средите на интервалите като представителни стойности:

$$S = \sqrt{\frac{\sum (X_u - \bar{X})^2 f}{n - 1}} = \sqrt{\frac{396{,}67}{29}} = 3{,}70 \text{ месеца.}$$

Стъпка 3. Стандартна грешка на средната.

$$SE_{\bar{X}} = \frac{S}{\sqrt{n}} = \frac{3{,}70}{\sqrt{30}} = 0{,}68 \text{ месеца.}$$

Стъпка 4. Критична стойност. Тъй като σ е неизвестно и е заместено с S, използваме $t_{0{,}975;29} = 2{,}045$.

Стъпка 5. Доверителен интервал. Изчисляваме с незакръглените междинни стойности и закръгляме едва крайния резултат:

$$95\ \%\ CI = 8{,}33 \pm 2{,}045 \cdot 0{,}675 = 8{,}33 \pm 1{,}38 = [6{,}95;\ 9{,}71] \text{ месеца.}$$

Заключение. Оценената популационна средна преживяемост е 8,33 месеца, а нейният 95 % доверителен интервал е от **6,95 до 9,71 месеца**. Ако многократно вземаме извадки по същия начин и строим интервали със същата процедура, около 95 % от тях ще съдържат истинската средна μ.

Не закръгляйте междинните резултати твърде рано

Ако заменим $8{,}33$ с 8 и $0{,}675$ с 0,7 още преди крайното изчисление, центърът и границите на интервала се изместват. Запазвайте поне три–четири знака в междинните стъпки и закръгляйте крайния отговор според точността на измерването.

3.4 Доверителен интервал за относителен дял

Когато $\hat{p}$ е записан като процент, стандартната грешка на относителния дял се изчислява с:

$$SE_{\hat{p}} = S_p = \sqrt{\frac{\hat{p}(100 - \hat{p})}{n}}.$$

При достатъчно голяма извадка 95 % интервалът по нормалното приближение е:

$$CI = \hat{p} \pm 1{,}96 \cdot SE_{\hat{p}}.$$

Приближението е приемливо, когато броят на наблюденията със събитието и броят без събитието не са малки; практично правило е и двата да са поне 5. При редки събития или малки извадки се предпочита интервал на Уилсън или точен биномиален интервал, изчислен със статистически софтуер. Интервалът по горната проста формула може иначе да даде невъзможни граници под 0 % или над 100 %.

3.4.1 Решен пример: доверителен интервал за дял

Условие

Вид диета	С подобрение	Без подобрение	Общо
Безглутенова	56	44	100
Контролна	45	55	100
Общо	**101**	**99**	**200**

Задача. Да се изчисли 95 % доверителен интервал за популационния дял с подобрение при безглутенова диета P_1.

Стъпка 1. Извадков относителен дял.

$$\hat{p}_1 = \frac{56}{100} \cdot 100 = 56\,\%.$$

Стъпка 2. Стандартна грешка на относителния дял.

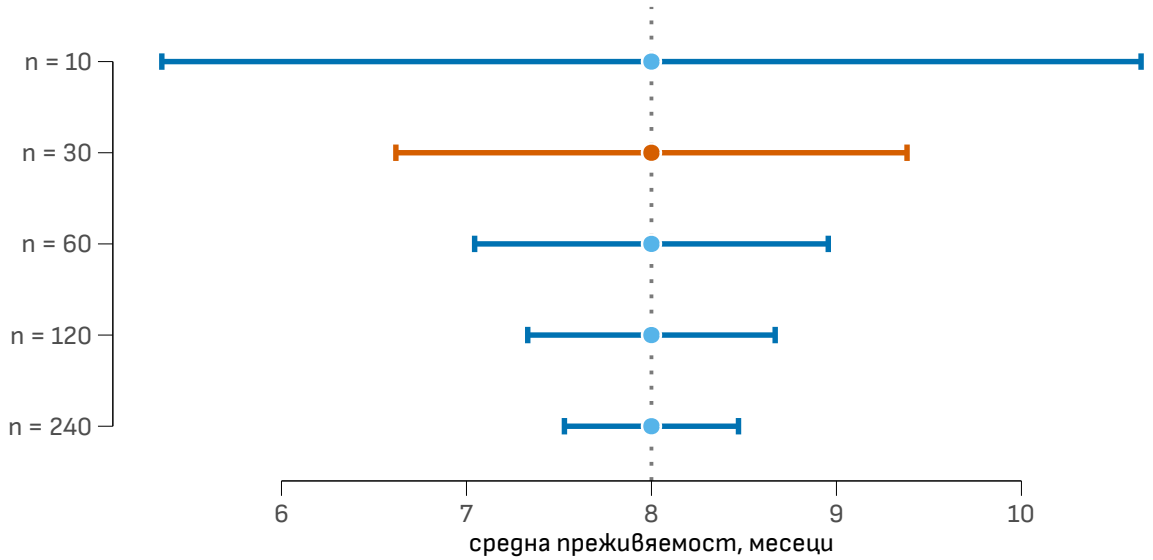
$$SE_{\hat{p}_1} = \sqrt{\frac{56(100-56)}{100}} = \sqrt{24{,}64} = 4{,}96 \text{ процентни пункта.}$$

Стъпка 3. Доверителен интервал.

$$95\ \%\ CI = 56 \pm 1{,}96 \cdot 4{,}96 = 56 \pm 9{,}73 = [46{,}27;\ 65{,}73]\ \%.$$

Заключение. Извадковият дял с подобрение е 56 %, а неговият 95 % доверителен интервал е от **46,27 % до 65,73 %**. Условията за нормалното приближение са изпълнени, защото има 56 наблюдения с подобрение и 44 без подобрение.

3.5 Какво променя ширината на интервала



Фиг. 3.3: Влияние на обема на извадката върху ширината на 95 % доверителния интервал. Средната (8 месеца) и стандартното отклонение (3,7 месеца) са еднакви; променя се само n. Червеният ред е примерът с n = 30.

Таблица 3.3: От какво зависи ширината на доверителния интервал

Ако...	ширината на интервала...	защото...
увеличим обема на извадката n	**намалява**	стандартната грешка намалява
увеличим стандартното отклонение S	**нараства**	стандартната грешка нараства
изберем 99 % вместо 95 %	**нараства**	критичната стойност е по-голяма

Ако...	ширината на интервала...	защото...
изберем 90 % вместо 95 %	**намалява**	критичната стойност е по-малка

> ⚠ Компромис между покритие и точност
>
> При едни и същи данни 99 % интервалът има по-високо покритие, но е по-широк и дава по-малко прецизна оценка. По-голяма прецизност без намаляване на нивото на доверителност се постига чрез по-голяма извадка или по-прецизни измервания, които намаляват разсейването.

3.6 Сравнение на две групи чрез доверителни интервали

Доверителните интервали са полезни за визуална ориентация, но отделните интервали на две групи не са доверителен интервал на **разликата между групите**.

> ! Визуална ориентация, а не формален тест
>
> При две независими групи липсата на препокриване между техните 95 % доверителни интервали е силно указание за разлика. Препокриването обаче не доказва, че разлика няма. Категоричният извод трябва да се основава на подходящ тест на хипотеза или на доверителен интервал за самата разлика (Секция 4).

> 💡 Пример
>
> 95 % CI за средното намаление на теглото при диета А е [4,5; 7,2] кг, а при диета Б — [2,2; 4,2] кг. Интервалите не се препокриват и данните убедително насочват към по-голямо средно намаление при диета А. За формално заключение се изчислява интервал на разликата или се прилага подходящ тест.
> Ако интервалите са [4,5; 7,2] и [2,2; 5,0], те се препокриват. Само от това не може да се заключи нито „има разлика“, нито „няма разлика“.

3.7 Чести грешки

Таблица 3.4: Чести грешки при доверителните интервали

#	Грешка	Как да я избегнете
1	„Има 95 % вероятност истинската стойност да е в този интервал“	Нивото описва дългосрочното покритие на метода
2	„95 % от пациентите имат стойност в интервала“	Доверителният интервал се отнася до параметър, а не до отделни пациенти

#	Грешка	Как да я избегнете
3	Използване на S вместо $SE_{\bar{X}}$ във формулата	Интервалът се строи със стандартната грешка
4	Използване на Z при малка извадка и неизвестно σ	За средна използвайте критична стойност от t-разпределението
5	Смесване на 0,56 и 56 във формулата за дял	Формулата с $(100 - \hat{p})$ изисква процент
6	Извод „няма разлика“ при препокриващи се интервали	Необходим е тест или интервал на разликата
7	Ранно закръгляване и пропусната мерна единица	Закръглете накрая и запишете единицата при двете граници

3.8 Задачи за самостоятелна работа и домашна работа

Задача 3.1

Изчислете 95 % доверителен интервал за средния популационен инкубационен период при пациенти с хепатит А. Използвайте t-критична стойност и интерпретирайте резултата с мерна единица.

Инкубационен период, дни	1–15	16–30	31–45	46–60	61–75	Общо
Честота f, брой пациенти	40	120	75	14	1	250

Задача 3.2

Изчислете 95 % доверителен интервал за средната популационна преживяемост при пациенти с белодробен карцином, лекувани с експериментална терапия.

Преживяемост, месеци	1–5	6–10	11–15	16–20	21–25	Общо
Честота f, брой пациенти	35	65	95	70	35	300

Задача 3.3

Използвайте данните от примера с безглутеновата диета (56 от 100 с подобрение), но изчислете **99 %** доверителен интервал. Как се променят точковата оценка и ширината на интервала спрямо 95 % CI?

Задача 3.4

В проспективно проучване на пациенти с миастения гравис тежки миастенни кризи са наблюдавани при 23 от 75 пациенти: $\hat{p} = 30{,}7\ \%$ и 95 % CI [20,1 %; 40,3 %]. Кое твърдение е най-коректно?
а) В популацията честотата непременно е по-голяма от 30,7 %.

б) Всеки отделен пациент има точно 30,7 % вероятност да развие тежка криза.
в) Има 95 % вероятност извадковият дял да бъде между 20,1 % и 40,3 %.
г) Интервалът е получен с метод, който при многократно повторение би покрил истинския популационен дял в около 95 % от случаите.

Задача 3.5

При еднакви S и ниво на доверителност изследовател увеличава извадката от 50 на 200 наблюдения. Колко пъти се променя стандартната грешка? Как това влияе върху ширината на интервала?

Задача 3.6

В пилотно проучване подобрение е наблюдавано при 2 от 20 пациенти. Проверете условието за нормалното приближение. Подходящо ли е да се използва простият интервал $\hat{p} \pm 1{,}96SE_{\hat{p}}$? Посочете по-подходящ избор.

Задача 3.7

Две независими групи имат 95 % доверителни интервали [21 %; 38 %] и [34 %; 48 %]. Какво може и какво не може да се заключи само от препокриването им?

3.9 Кратък тест

За всеки въпрос изберете **един** верен отговор. Ключът е в Секция Г.3.

1. Кое е параметър?

   а) средният пулс на 40 изследвани студенти;
   б) делът с подобрение сред 100 включени пациенти;
   в) истинският среден пулс на всички студенти от II курс;
   г) стандартното отклонение в конкретна извадка.

2. Ако обемът на извадката се увеличи четири пъти при непроменено S, стандартната грешка:

   а) се увеличава четири пъти;
   б) се увеличава два пъти;
   в) намалява два пъти;
   г) не се променя.

3. Коя интерпретация на 95 % доверителен интервал е правилна?

   а) 95 % от наблюденията са между двете граници;
   б) методът ще даде интервали, съдържащи истинския параметър, в около 95 % от многократните повторения;
   в) има 95 % вероятност извадковата средна да е в интервала;
   г) параметърът се изменя случайно между двете граници.

4. Кое действие разширява доверителния интервал при едни и същи данни?

   а) увеличаване на n;
   б) преминаване от 95 % към 90 % доверителност;
   в) преминаване от 95 % към 99 % доверителност;
   г) намаляване на стандартното отклонение.

5. При малка извадка и неизвестно популационно стандартно отклонение интервалът за средна използва:

   а) t-критична стойност с $n - 1$ степени на свобода;
   б) винаги $Z = 1{,}96$;
   в) стандартното отклонение без деление на $\sqrt{n}$;
   г) само извадковата средна.

6. Два 95 % доверителни интервала се препокриват. Кой извод е допустим?

   а) групите със сигурност не се различават;
   б) групите със сигурност се различават;
   в) само препокриването не е достатъчно; нужен е интервал на разликата или подходящ тест;
   г) и двете средни са равни на нула.

4 Проверка на хипотези. Z-тест и t-тест

Цел на упражнението

След това упражнение трябва да умеете:

1. да формулирате нулева и алтернативна хипотеза за конкретна задача;
2. да обяснявате какво е p-стойност и какво **не** е тя;
3. да различавате грешка от I и от II род и да свързвате мощността с тях;
4. да избирате подходящ статистически тест според вида на признака, броя групи и дизайна;
5. да прилагате Z-тест за два относителни дяла и t-тест за две независими средни и да формулирате извода.

Какво трябва да знаете отпреди

От Секция 2: средна аритметична, стандартно отклонение, относителен дял. От Секция 3: стандартна грешка и доверителен интервал.

4.1 Хипотези

Статистическата хипотеза е проверимо твърдение за параметър или за връзка в генералната съвкупност. Извадковите данни се използват, за да се оцени дали са съвместими с това твърдение. При разглежданите в тази глава тестове процедурата започва с допускането, че **нулевата хипотеза** е вярна.

Определения

Нулевата хипотеза H_0 при тестовете за разлика твърди, че популационната разлика е нула. Например $H_0 : \mu_1 - \mu_2 = 0$ или $H_0 : P_1 - P_2 = 0$.
Алтернативната хипотеза H_1 твърди, че популационната разлика не е нула. При двустранен тест: $H_1 : \mu_1 - \mu_2 \neq 0$.
Статистическият тест не изисква изследователят да вярва в H_0. Тя се приема временно като изходно допускане, за да се изчисли колко необичайни са наблюдаваните данни при липса на популационна разлика.

Пример

Тества се нов медикамент за хипертония. Едната група получава медикамента, другата — плацебо. В края на изследването кръвното налягане в експерименталната група е средно 10 mmHg по-ниско.
H_0: Популационната средна промяна в кръвното налягане е еднаква при медикамента и плацебото: $\mu_1 - \mu_2 = 0$.

H_1: Популационната средна промяна е различна при двете групи: $\mu_1 - \mu_2 \neq 0$. Разликата от 10 mmHg е извадкова оценка. Тестът показва доколко тя е съвместима с H_0, но сам по себе си не доказва причинен ефект. Причинният извод зависи и от дизайна, рандомизацията, придържането към лечението и възможните систематични грешки.

4.2 p-стойност

Определение

p-стойността е вероятността, **ако H_0 и предпоставките на теста са верни**, да се получи тестова статистика поне толкова отдалечена от очакваното при H_0, колкото наблюдаваната.

p е число между 0 и 1. Малката p-стойност означава, че данните са слабо съвместими с H_0. Голямата p-стойност означава, че данните не дават достатъчно основание за отхвърляне на H_0; тя не показва, че H_0 е вероятно вярна.

Нивото на значимост α е предварително избран праг за решението. В медицинските изследвания често се използва $\alpha = 0{,}05$. Прагът трябва да се зададе преди анализа, а не да се променя след получаването на резултатите.

Правило за решението

1. При $p < \alpha$ **отхвърляме** H_0. Резултатът е статистически значим при избраното α.
2. При $p \geq \alpha$ **не отхвърляме** H_0. Данните не са достатъчни за статистически значим извод.

В част от учебните материали второто решение е записано като „приемаме H_0“. Поточната формулировка е „не отхвърляме“, защото незначимият резултат не доказва равенство.

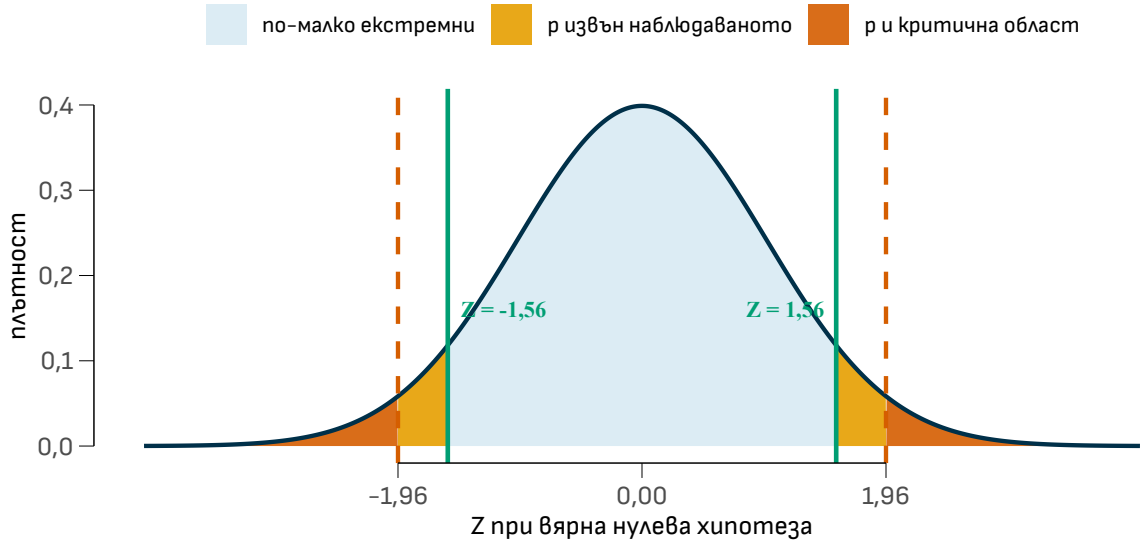
Как да се чете $p = 0{,}09$

Ако H_0 и предпоставките на теста са верни, при многократно повторение на същия дизайн в около 9 % от опитите ще се получи тестова статистика поне толкова екстремна, колкото наблюдаваната. Понеже $0{,}09 > 0{,}05$, не отхвърляме H_0. Това не означава, че вероятността H_0 да е вярна е 9 %.

Какво p-стойността НЕ е

- p **не е** вероятността нулевата хипотеза да е вярна.
- p **не е** вероятността резултатът да се повтори при ново проучване.
- p **не измерва** големината на ефекта. Много малко p при огромна извадка може да съпровожда клинично незначим ефект.

- $p \geq 0{,}05$ **не доказва**, че групите са еднакви или че ефект няма.



Фиг. 4.1: Двустранна p-стойност при Z = 1,56. Оранжевите и червените опашки заедно съдържат стойностите поне толкова екстремни, колкото наблюдаваната; общата им площ е p = 0,12. Прекъснатите линии са критичните граници ±1,96 при α = 0,05.

4.3 Грешки от I и от II род

При всяко решение, взето въз основа на извадка, са възможни два вида грешки (Таблица 4.1).

Таблица 4.1: Видове грешки при проверката на хипотеза

	H_0 е **вярна**	H_0 е **грешна**
Отхвърляме H_0	грешка от I род α	правилно решение
Не отхвърляме H_0	правилно решение	грешка от II род β

Грешката от I род α означава да обявим за „научно откритие“ нещо, което в действителност не съществува. Предварително зададеното α ограничава вероятността за такава грешка при вярна H_0.

Грешката от II род β означава да пропуснем действително съществуваща разлика. Тя зависи от размера на ефекта, вариабилността, обема на извадката, избраното α и статистическия тест.

! Двете грешки са свързани

При непроменени обем, ефект и тест намаляването на α от 0,05 на 0,01 намалява риска от грешка от I род, но **увеличава** риска от грешка от II род. Едновременното намаляване на двата риска обикновено изисква по-голяма извадка или по-прецизни измервания.

4.4 Мощност

Мощността е вероятността тестът да отхвърли H_0, когато предварително зададена реална разлика действително съществува:

$$\text{мощност} = 1 - \beta$$

Тя зависи от четири фактора:

1. **Ниво на значимост** α. Увеличаването на α (от 0,01 на 0,05) намалява β и увеличава мощността.
2. **Размер на ефекта** Δ. Колкото по-голяма е действителната разлика, толкова по-лесно се доказва.
3. **Вариабилност**. Колкото по-малко варират наблюденията, толкова по-малка е стандартната грешка и толкова по-голяма е мощността.
4. **Обем на извадките** n. Повече наблюдения → по-малка стандартна грешка → по-висока мощност.

> Типов изпитен въпрос
>
> *При равни други условия най-малък брой единици на наблюдение ще е необходим при:* еднородна популация и голям по размер клиничен ефект. Хомогенността намалява разсейването, а големият ефект се открива по-лесно - и двете увеличават мощността, така че стигат по-малко наблюдения.

4.5 Двустранна и едностранна алтернатива

При **двустранен тест** алтернативната хипотеза допуска разлика и в двете посоки: $H_1 : \mu_1 - \mu_2 \neq 0$. При **едностранен тест** посоката е зададена предварително, например $H_1 : \mu_1 - \mu_2 > 0$. Едностранен тест не се избира след като посоката на резултата вече е известна. Всички решени примери в тази глава са двустранни.

4.6 Избор на статистически тест

Изборът започва с вида на резултата: количествен, ординален или категориален. След това се определят броят на групите и дали наблюденията са независими или свързани. Накрая се проверяват предпоставките на конкретния тест.

Таблица 4.2: Ориентир за избор на статистически тест

Резултат и дизайн	Една група	Две групи	Три или повече групи
количествен, независим	едногрупов t-тест	t-тест на Welch	ANOVA
количествен, свързан	–	t-тест за свързани извадки	ANOVA с повторни измервания
количествен/ординален, непараметричен	знаково-рангов тест на Wilcoxon	Mann–Whitney за независими; Wilcoxon за свързани	Kruskal–Wallis за независими; Friedman за свързани
категориален, независим	биномиален тест или Z-тест за един дял	Z-тест за два дяла или χ^2-тест	χ^2-тест
дихотомен, свързан	–	тест на McNemar	Q-тест на Cochran

„Непараметричен“ не означава автоматично „тест за същата средна без нормалност“. Например тестът на Mann–Whitney сравнява разпределенията и при допълнителни предположки може да се тълкува като разлика в местоположението. При две независими средни t-тестът на Welch често е устойчив при умерено отклонение от нормалността, особено при достатъчно големи групи и липса на крайни стойности.

Как да разпознаете зависим дизайн

Питайте се: *измерена ли е една и съща единица на наблюдение повече от веднъж?* „Изменение спрямо изходните стойности“, „преди и след лечение“, „ляво и дясно око на един пациент“ – всичко това е зависим дизайн.

4.7 Z-тест за сравнение на два относителни дяла

4.7.1 Решен пример

Условие

Режим с безглутенова диета	С подобрение	Без подобрение	Общо
Да	56	44	100
Не	45	55	100
Общо	**101**	**99**	**200**

$\hat{p}_1 = 56\,\%$, $\hat{p}_2 = 45\,\%$, разлика (клиничен ефект) $\Delta = 11$ процентни пункта.
Задача. Има ли разлика между двете групи по отношение на относителния дял на случаите с подобрение?

Стъпка 1. Хипотези.

$H_0 : P_1 - P_2 = 0$ - популационните дялове с подобрение са еднакви.

$H_1 : P_1 - P_2 \neq 0$ - популационните дялове с подобрение са различни.

Тестът е двустранен, а предварително избраното ниво е $\alpha = 0{,}05$.

Стъпка 2. Изчисляване на тестовата статистика.

Първо се изчислява **обединеният** относителен дял:

$$\hat{p} = \frac{m_1 + m_2}{n_1 + n_2} \times 100 = \frac{56 + 45}{100 + 100} \times 100 = \frac{101}{200} \times 100 = 50{,}5\,\%$$

След това:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p} \cdot (100 - \hat{p}) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$Z = \frac{56 - 45}{\sqrt{50{,}5 \cdot 49{,}5 \cdot \left(\frac{1}{100} + \frac{1}{100}\right)}} = \frac{11}{\sqrt{49{,}995}} = \frac{11}{7{,}07} = 1{,}56$$

Стъпка 3. Определяне на p-стойността. Сравняваме абсолютната стойност $|Z|$ с двустранните критични стойности на нормалното разпределение (Таблица 4.3).

Таблица 4.3: Критични стойности на Z

$p = 0{,}10$	$p = 0{,}05$	$p = 0{,}01$
$Z = 1{,}645$	$Z = 1{,}960$	$Z = 2{,}576$

Изчисленото $|Z| = 1{,}56$ е по-малко от 1,645. По таблицата записваме $p > 0{,}10$; статистическият софтуер дава двустранно $p = 0{,}120$.

Стъпка 4. Заключение. Понеже $p = 0{,}120 > \alpha$, не отхвърляме H_0. Данните не дават достатъчно доказателства за разлика между популационните дялове с подобрение. Това не доказва, че двете диети са еднакво ефективни.

Стъпка 5. Размер и прецизност на ефекта. Извадковата разлика е $56\% - 45\% = 11$ процентни пункта. За доверителния интервал на разликата използваме необединена стандартна грешка:

$$SE_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{56 \cdot 44}{100} + \frac{45 \cdot 55}{100}} = 7{,}03 \text{ процентни пункта.}$$

$$95\,\%\, CI = 11 \pm 1{,}96 \cdot 7{,}03 = [-2{,}8;\ 24{,}8] \text{ процентни пункта.}$$

Интервалът включва нулата и води до същия статистически извод. Той показва и широкия диапазон от популационни разлики, съвместими с данните.

Защо стандартните грешки са различни

При Z-теста обединеният дял се използва, защото изчислението се извършва при H_0 : $P_1 = P_2$. При доверителния интервал не налагаме равенство и използваме отделните извадкови дялове. Не смесвайте двете стандартни грешки.

4.8 t-тест за сравнение на две независими извадки

t-тестът на Welch проверява дали две независими популационни средни се различават. Той не изисква еднакви дисперсии и затова е подходящ основен избор при две независими групи.

Условия за прилагане

1. Резултатът е количествен, а наблюденията в двете групи са независими.
2. Извадките са случайни или дизайнът позволява обобщение към целевата популация.
3. При малки извадки разпределенията са приблизително нормални и без крайни стойности. При по-големи извадки тестът е устойчив на умерена асиметрия.
4. Наблюденията вътре във всяка група са независими. Повторни измервания при един пациент изискват тест за свързани извадки или друг подход.

4.8.1 Решен пример

Условие

Вид лечение	Обем на извадката	Средна преживяемост	Стандартно отклонение
Експериментално	$n_1 = 30$	$\bar{X}_1 = 15$ месеца	$S_1 = 4{,}5$ месеца
Химиотерапия	$n_2 = 30$	$\bar{X}_2 = 11$ месеца	$S_2 = 1{,}5$ месеца

Разлика (клиничен ефект) $\Delta = 15 - 11 = 4$ месеца.
Задача. Има ли разлика между двете групи по отношение на средната преживяемост?

Стъпка 1. Хипотези.

$H_0 : \mu_1 - \mu_2 = 0$ - популационните средни преживяемости са еднакви.

$H_1 : \mu_1 - \mu_2 \neq 0$ - популационните средни преживяемости са различни.

Тестът е двустранен и $\alpha = 0{,}05$.

Стъпка 2. Изчисляване на тестовата статистика.

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$t = \frac{|15 - 11|}{\sqrt{\frac{4{,}5^2}{30} + \frac{1{,}5^2}{30}}} = \frac{4}{\sqrt{0{,}75}} = \frac{4}{0{,}866} = 4{,}62.$$

ℹ Същата формула, записана чрез средните грешки

Тъй като $S_{\bar{X}} = S/\sqrt{n}$, то $S_{\bar{X}}^2 = S^2/n$. Затова формулата може да се запише и така:

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{S_{\bar{X}_1}^2 + S_{\bar{X}_2}^2}}$$

Двата записа са **един и същ** израз. Използвайте втория, когато в условието са дадени направо средните грешки (записът $\bar{X} \pm S_{\bar{X}}$), и първия, когато са дадени стандартните отклонения и обемите.

Стъпка 3. Определяне на p-стойността.

При теста на Welch степените на свобода се изчисляват с приближението на Welch–Satterthwaite:

$$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(S_2^2/n_2\right)^2}{n_2 - 1}} = 35{,}37.$$

За ръчно търсене в таблицата (Секция В) използваме близкия по-малък цял ред, например $df = 35$:

df	$p = 0{,}10$	$p = 0{,}05$	$p = 0{,}01$
35	1,690	2,030	2,724

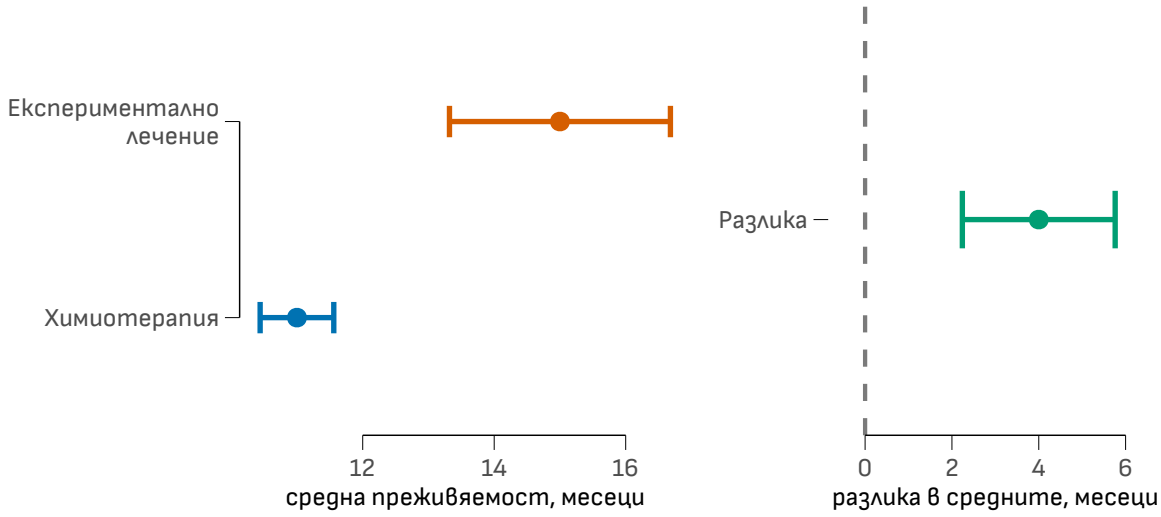
Изчисленото $t = 4{,}62$ е по-голямо от 2,724, следователно $p < 0{,}01$. Статистическият софтуер дава двустранно $p < 0{,}001$.

Стъпка 4. Заключение. Понеже $p < 0{,}001 < \alpha$, отхвърляме H_0. Данните показват статистически значима разлика в средната преживяемост между двете групи.

Стъпка 5. Размер и прецизност на ефекта. Оценената разлика е $15 - 11 = 4$ месеца. Нейният 95 % доверителен интервал е:

$$4 \pm t_{0{,}975;35{,}37} \cdot 0{,}866 = 4 \pm 2{,}03 \cdot 0{,}866 = [2{,}24;\ 5{,}76] \text{ месеца.}$$

Интервалът не включва нулата и е съгласуван с резултата от теста. Изводът, че лечението причинява по-добра преживяемост, е оправдан само ако дизайнът и изпълнението на проучването позволяват причинна интерпретация.



Фиг. 4.2: Средната преживяемост и 95 % доверителните интервали за двете групи (ляво) и оценената разлика между средните (дясно). Интервалът на разликата не включва нулата.

! Важни зависимости

При равни други условия:

- по-голяма **абсолютна разлика** между средните → по-голяма тестова статистика → по-малка p-стойност;
- по-ниска **вариабилност** (малко S) → по-малка стандартна грешка → по-голяма тестова статистика;
- по-голям **обем на извадките** → по-малка стандартна грешка → по-голяма тестова статистика.

Обобщено: **колкото по-голяма е изчислената тестова статистика, толкова по-малка е p-стойността.**

4.9 Задължителен изпитен въпрос

⚠ Формат на въпроса

Проучване достига до резултат $[Z;\ t,\ df;\ \chi^2,\ df]$. На колко е равна p-стойността в този случай и какво означава това? Какъв извод може да се направи?

Примерен въпрос и отговор

Въпрос. Проучване на нивото на затлъстяване при пушачи и непушачи достига до резултат $t = 3{,}45$ $(df = 50)$. На колко е равна p-стойността, какво означава това и какъв извод следва?
Отговор.

- *На колко е равна p-стойността?* При $df = 50$ критичната стойност за двустранно $p = 0{,}01$ е 2,678. Тъй като $3{,}45 > 2{,}678$, по таблицата следва $p < 0{,}01$; точната стойност е приблизително 0,001.
- *Какво означава това?* Ако H_0 и предпоставките на теста са верни, вероятността да се получи $|t| \geq 3{,}45$ е около 0,1 %.
- *Какъв извод следва?* Има статистически значима разлика в нивото на затлъстяване между пушачите и непушачите и отхвърляме H_0 при $\alpha = 0{,}05$. Без оценка и доверителен интервал не можем да определим посоката, големината или практическата важност на разликата.

4.10 Статистическа и клинична значимост

Статистически значимата разлика не е автоматично клинично значима, а статистически незначимият резултат не доказва липса на клинично важен ефект.

⚠ Определения

Статистическата значимост означава, че p е под предварително избраното α при използвания модел и неговите предпоставки.
Клиничната значимост се определя от големината на ефекта, неговия доверителен интервал, ползите, вредите, разходите и предварително зададената минимална клинично важна разлика.

Разлика от 2 mmHg в систолното налягане може да има много малка p-стойност при огромна извадка, но практическото ѝ значение трябва да се оцени в конкретния клиничен и популационен контекст. Обратно, клинично важен ефект може да остане статистически недоказан, ако извадката е твърде малка.

4.11 Чести грешки

Таблица 4.4: Чести грешки при проверката на хипотези

#	Грешка	Как да я избегнете
1	Формулиране на H_0 като „има разлика“	При разглежданите тестове H_0 задава нулева популационна разлика
2	„Доказахме, че разлика няма“ при $p > 0{,}05$	Не отхвърляме H_0; липсата на доказателство не е доказателство за липса
3	Тълкуване на p като вероятност H_0 да е вярна	p е вероятността за такива данни, **ако** H_0 е вярна
4	t-тест при зависими групи („преди и след“)	Проверете дизайна преди да изберете тест
5	t-тест при малка извадка със силна асиметрия и крайни стойности	Проверете данните и използвайте подходящ устойчив или непараметричен метод
6	Изчисляване на Z без обединения дял	Обединеният дял е задължителна стъпка
7	Извод, изчерпващ се с числото p	Изводът е изречение с имената на променливите

4.12 Задачи за самостоятелна работа и домашна работа

Задача 4.1

400 пациенти с множествена склероза са включени в клинично проучване. 25 от 200 лекувани с експериментална терапия рецидивират в рамките на 6 месеца след края на терапевтичния курс, докато при конвенционалната терапия този брой е 45 от 200. Има ли статистически значима разлика между двата вида терапия по отношение на относителния дял рецидивиращи пациенти?

Задача 4.2

600 пациенти с белодробен карцином са включени в клинично проучване. 250 от 300 лекувани с експериментална терапия преживяват поне 1 година, докато при конвенционалната терапия този брой е 190 от 300.
Има ли статистически значима разлика по отношение на относителния дял на пациентите, преживяващи поне 1 година?

Задача 4.3

Пушачи със затлъстяване са включени в клинично проучване на нов продукт за понижаване на холестерола. Три месеца след началото средната стойност (± стандартна грешка) е 213 ± 6 mg/dL за експерименталната група и 228 ± 7 mg/dL за контролната. Изчислете тестовата статистика. Може ли да се определи точната p-стойност, ако не са дадени обемите или степените на свобода?

Задача 4.4

Пациенти със стомашен аденокарцином са разделени на 3 групи според терапията. Сравняват се 4-седмичните изменения в туморния размер спрямо изходните стойности; величината не е нормално разпределена. Кой тест следва да се приложи?
а) H-тест на Kruskal–Wallis б) ANOVA в) тест на Friedman за свързани извадки г) ANOVA с повторни измервания

Задача 4.5

При равни други условия, ако намалим нивото на значимост от $\alpha = 0{,}05$ на $\alpha = 0{,}01$:
а) вероятността за грешка от I род ще нарасне, а за II род ще намалее; б) вероятността за грешка от II род ще нарасне, а за I род ще намалее; в) и двете вероятности ще нараснат; г) и двете вероятности ще намалеят.

Задача 4.6

Пациенти с белодробен карцином са разделени на 2 групи според терапията. Лекуваните с експериментална терапия преживяват средно 10,1 месеца, а лекуваните с химиотерапия — 8,7 месеца ($p = 0{,}08$). Кое от следните твърдения е вярно?
а) Приемаме нулевата хипотеза, че има статистически значима разлика. б) t-тестът е подходящ, защото средните са приблизително равни. в) Тъй като резултатът не е значим, може да се премине към t-тест за свързани извадки. г) Подходяща алтернативна хипотеза е, че двете групи имат различна преживяемост в зависимост от вида на лечението.

Задача 4.7

Проучване на нивото на физическа активност (часове тренировка седмично) при пушачи и непушачи достига до $t = 2{,}96$ при $df = 40$. Отговорете по схемата на задължителния изпитен въпрос.

Задача 4.8

Мощността на изследователския дизайн зависи от четири фактора. Дайте пример за комбинация от тях, която постига възможно най-голяма мощност.

Задача 4.9

При 40 пациенти концентрацията на C-реактивен протеин е измерена преди и след лечение. Разликите „след минус преди“ са приблизително нормално разпределени. Кой тест е подходящ и как се формулира H_0?

Задача 4.10

В голямо проучване разликата в средното систолно налягане е 1,8 mmHg, 95 % CI [1,2; 2,4] mmHg и $p < 0{,}001$. Разграничете статистическия извод от оценката на клиничната значимост.

4.13 Кратък тест

За всеки въпрос изберете **един** верен отговор. Ключът е в Секция Г.4.

1. Какво представлява p-стойността?

   а) вероятността H_0 да е вярна;
   б) вероятността при вярна H_0 да се получи поне толкова екстремна тестова статистика;
   в) вероятността резултатът да се повтори;
   г) големината на клиничния ефект.

2. При $p = 0{,}18$ и $\alpha = 0{,}05$:

   а) доказваме, че H_0 е вярна;
   б) отхвърляме H_0;
   в) не отхвърляме H_0;
   г) доказваме, че групите са еквивалентни.

3. Кое увеличава мощността при равни други условия?

   а) по-малка извадка;
   б) по-малък действителен ефект;
   в) по-голяма вариабилност;
   г) по-голяма извадка.

4. Кой тест е подходящ за нормално разпределена количествена величина, измерена преди и след лечение при едни и същи пациенти?

   а) t-тест за свързани извадки;
   б) t-тест на Welch за независими извадки;
   в) тест на Mann–Whitney;
   г) Z-тест за два дяла.

5. Намаляването на α от 0,05 на 0,01 при непроменени обем и ефект:

   а) увеличава грешката от I род и намалява грешката от II род;
   б) намалява грешката от I род и увеличава грешката от II род;

в) намалява и двете грешки;
г) не променя мощността.

6. Кое трябва да се представи заедно с p-стойността?

а) само броят на значимите тестове;
б) оценка на ефекта и доверителен интервал;
в) само избраното α;
г) единствено посоката на алтернативната хипотеза.

5 Връзка между качествени признаци. χ^2-тест

Цел на упражнението

След това упражнение трябва да умеете:

1. да разпознавате кога се прилага χ^2-тест и да го отличавате от t-теста и от корелационния анализ;
2. да построявате таблица на съчетаемост от словесно условие и да намирате маргиналните суми;
3. да изчислявате очакваните честоти при вярна нулева хипотеза и да проверявате условията за прилагане на теста;
4. да изчислявате χ^2-статистиката и степените на свобода и да определяте p-стойността по таблица;
5. да оценявате силата на връзката чрез коефициента φ или V на Cramér и да формулирате пълен извод.

Какво трябва да знаете отпреди

От Секция 2: качествени признаци, номинална и ординална скала, относителен дял. От Секция 4: нулева и алтернативна хипотеза, p-стойност, ниво на значимост, степени на свобода.

5.1 Кога се използва χ^2-тестът

Изборът на статистически метод се определя от вида на **двете** променливи, чиято връзка изследваме (Таблица 5.1).

Таблица 5.1: Избор на метод според вида на двете променливи

Първа променлива	Втора променлива	Подходящ метод
количествена	качествена с 2 категории	t-тест (Секция 4)
качествена с 2 категории	качествена с 2 категории	Z-тест за два дяла или χ^2-тест
качествена	качествена (произволен брой категории)	**χ^2-тест** (тази глава)
количествена	количествена	корелация и регресия (Секция 6, Секция 7)

Когато **и двете** променливи са качествени, средна аритметична не може да се изчисли и корелационен коефициент на Pearson не е приложим. Разполагаме само с **броя случаи** във всяка комбинация от категории. Точно върху такива данни работи **тестът на Pearson за независимост**, известен още като χ^2**-тест за асоциация**.

Типични въпроси, на които той отговаря:

- Има ли зависимост между честотата на физиотерапията и степента на възстановяване?
- Има ли зависимост между вида на антибиотичния режим и успеха на ерадикацията?
- Има ли зависимост между продължителността на трудовия стаж и наличието на неврологични оплаквания?

i Връзката между Z-теста и χ^2-теста

При таблица 2 × 2 двустранният Z-тест за два относителни дяла и χ^2-тестът водят до един и същ извод, защото е изпълнено $\chi^2 = Z^2$. Например $Z = 1{,}56$ съответства на $\chi^2 = 1{,}56^2 = 2{,}43$ при $df = 1$, и двете стойности дават $p > 0{,}10$. χ^2-тестът е по-общ: той работи и при повече от две категории, където Z-тестът не е приложим.

5.2 Таблица на съчетаемост

! Определение

Таблицата на съчетаемост (контингентна таблица) разпределя единиците на наблюдение едновременно по две качествени променливи. В клетките ѝ стоят **наблюдаваните честоти** O, а по краищата — сумите по редове R_i, сумите по колони C_j и общата сума N, наричани **маргинални суми**.

Общият вид на таблица с R реда и C колони е представен в Таблица 5.2.

Таблица 5.2: Общ вид на таблица на съчетаемост

	Категория B_1	Категория B_2	Сума по ред
Категория A_1	O_{11}	O_{12}	R_1
Категория A_2	O_{21}	O_{22}	R_2
Сума по колона	C_1	C_2	N

! Три правила при построяването на таблицата

1. Всяка единица на наблюдение попада в **точно една** клетка. Сборът на всички клетки е равен на общия брой изследвани лица.
2. Категориите на всяка променлива са **изчерпателни** и **взаимно изключващи се**.
3. В клетките стоят **броеве**, а не проценти. Процентите се изчисляват допълнително, само за словесното описание на таблицата.

5.3 Логиката на теста

χ^2-тестът сравнява две картини на едни и същи данни:

- **Наблюдаваните честоти** O — това, което действително е преброено. Това са **фактите**.
- **Очакваните честоти** E — това, което бихме преброили, ако между двете променливи **няма никаква връзка**. Това са **очакванията при вярна нулева хипотеза**.

Ако двете картини съвпадат, няма основание да се твърди, че между променливите има зависимост. Колкото по-голямо е несъответствието между тях, толкова по-силно е доказателството срещу нулевата хипотеза.

> Интуицията зад очакваните честоти
>
> Ако честотата на физиотерапията няма никакво значение за изхода, делът на напълно възстановените трябва да бъде **един и същ** при всеки режим и да съвпада с общия дял в цялата извадка. Очакваните честоти са точно този общ дял, приложен поотделно към всеки ред.

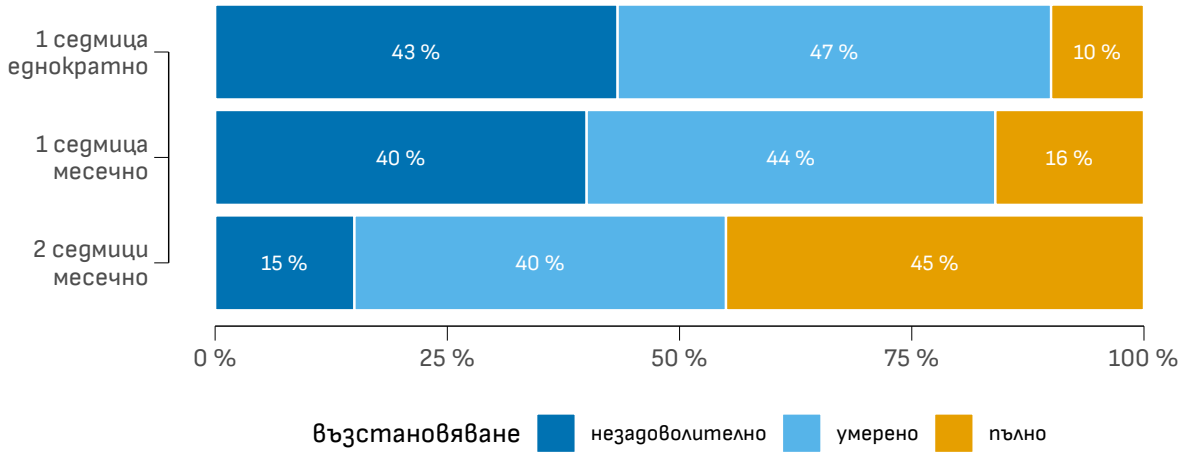
5.4 Решен пример: таблица 3 × 3

> Условие
>
> Проследени са 500 пациенти, получаващи различен режим физиотерапия след ставно протезиране. Записана е постигнатата степен на възстановяване.
>
> | Честота на физиотерапия | Незадоволително | Умерено | Пълно | Всичко |
> |---|---|---|---|---|
> | 1 седмица еднократно | 65 | 70 | 15 | 150 |
> | 1 седмица месечно | 100 | 110 | 40 | 250 |
> | 2 седмици месечно | 15 | 40 | 45 | 100 |
> | **Всичко** | **180** | **220** | **100** | **500** |
>
> **Задача.** Има ли връзка (зависимост) между честотата на физиотерапевтичното лечение и степента на възстановяване?

Преди изчисленията данните се описват словесно. Делът на напълно възстановените е $15/150 = 10{,}0\ \%$ при най-слабия режим, $40/250 = 16{,}0\ \%$ при средния и $45/100 = 45{,}0\ \%$ при най-интензивния (Фиг. 5.1). Разликите изглеждат големи, но за извод е необходим тест.

5.4.1 Стъпка 1. Хипотези и ниво на значимост

H_0: Няма връзка (зависимост) между честотата на физиотерапевтичното лечение и степента на възстановяване. Двете променливи са независими и честотата на физиотерапи-



Фиг. 5.1: Разпределение на изхода при трите режима физиотерапия. Делът на напълно възстановените нараства от 10 % при най-слабия до 45 % при най-интензивния режим.

ята не е фактор за изхода.

H_1: Има връзка между честотата на физиотерапевтичното лечение и степента на възстановяване. Двете променливи не са независими.

Работна хипотеза е нулевата; предварително избраното ниво на значимост е $\alpha = 0{,}05$.

> **Защо алтернативата няма посока**
>
> χ^2-тестът не проверява дали един конкретен режим е по-добър от друг. Той проверява само дали разпределението на изхода е **едно и също** при всички режими. Затова алтернативната хипотеза не съдържа посока и тестът няма едностранен вариант в смисъла от Секция 4.

5.4.2 Стъпка 2. Очаквани честоти

$$E_{ij} = \frac{R_i \times C_j}{N}$$

където R_i е сумата по i-тия ред, C_j — сумата по j-тата колона, а N — общият брой наблюдения.

За клетката „1 седмица еднократно × незадоволително“:

$$E_{11} = \frac{150 \times 180}{500} = \frac{27\,000}{500} = 54{,}0$$

За клетката „2 седмици месечно × пълно“:

$$E_{33} = \frac{100 \times 100}{500} = 20{,}0$$

Пълният набор очаквани честоти е даден в Таблица 5.3.

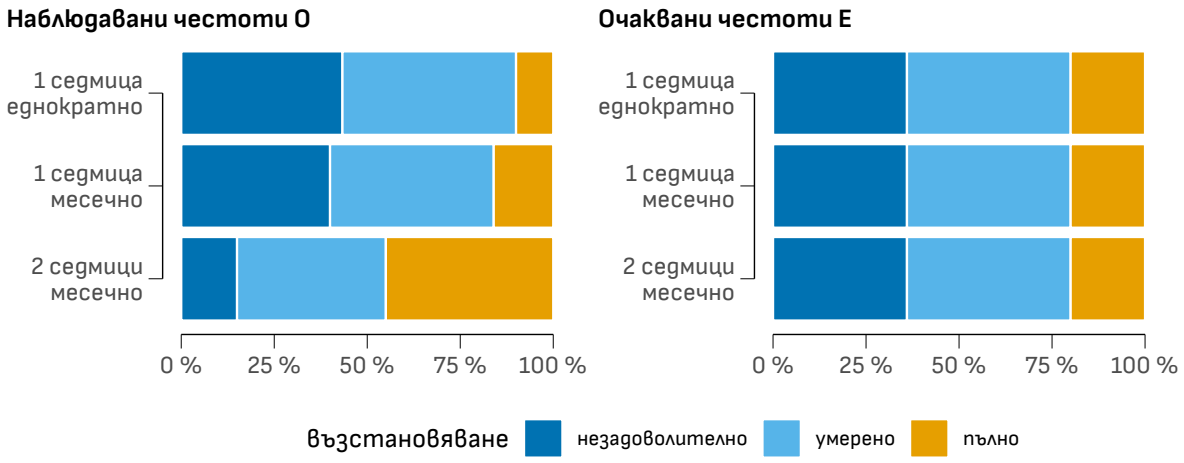
Таблица 5.3: Очаквани (теоретични) честоти при вярна H_0

Честота на физиотерапия	Незадоволи-телно	Умерено	Пълно	Всичко
1 седмица еднократно	54,0	66,0	30,0	150
1 седмица месечно	90,0	110,0	50,0	250
2 седмици месечно	36,0	44,0	20,0	100
Всичко	**180**	**220**	**100**	**500**

Две проверки и една икономия

- Маргиналните суми на очакваните честоти **съвпадат** с тези на наблюдаваните. Ако при вас не съвпадат, има аритметична грешка.
- Затова последната клетка във всеки ред и всяка колона може да се получи чрез изваждане: $150 - 54{,}0 - 66{,}0 = 30{,}0$.
- Забележете, че при вярна H_0 делът на напълно възстановените е еднакъв във всички редове: $30/150 = 50/250 = 20/100 = 20\ \%$. Точно това означава „няма връзка“.

Съпоставката между двете картини е представена на Фиг. 5.2.



Фиг. 5.2: Наблюдавани честоти (вляво) и очаквани честоти при вярна нулева хипотеза (вдясно). При независимост разпределението на изхода би било еднакво и в трите режима; наблюдаваните данни се отклоняват от тази картина.

5.4.3 Стъпка 3. Проверка на условията

> **!** Условия за прилагане на χ^2-теста
>
> 1. Наблюденията са **независими** — всяко лице попада само в една клетка.
> 2. В клетките стоят **броеве**, а не проценти или средни стойности.
> 3. Поне **80 %** от очакваните честоти са ≥ 5.
> 4. **Няма** очаквана честота, по-малка от 1.
>
> Ако условие 3 или 4 не е изпълнено, при таблица 2 × 2 се използва точният тест на Fisher, а при по-големи таблици се обединяват съседни категории. Обединяването се решава **преди** да се види резултатът.

В примера най-малката очаквана честота е 20,0. И четирите условия са изпълнени.

5.4.4 Стъпка 4. Изчисляване на χ^2-статистиката

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

където сумирането е по всички клетки на таблицата, а за всяка клетка:

- $O - E$ е разликата между факта и очакването;
- $(O - E)^2$ е квадратът на разликата — квадратирането премахва знаците, така че положителните и отрицателните отклонения да не се занулят взаимно;
- делението на E претегля разликата спрямо мащаба на клетката: разлика от 10 случая е много при очаквани 20 и малко при очаквани 500.

Таблица 5.4: Изчисляване на χ^2 по клетки

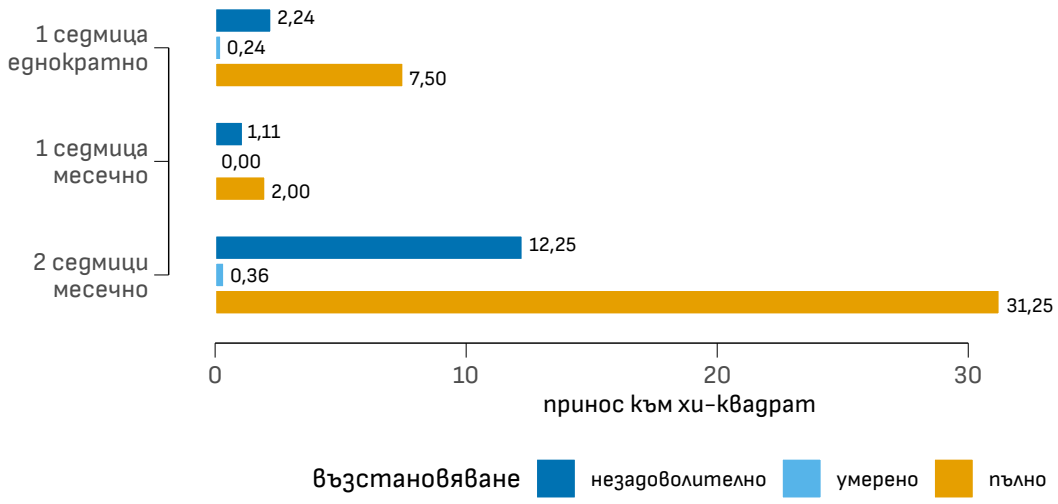
Клетка	O	E	$O - E$	$(O - E)^2$	$(O - E)^2/E$
еднократно × незадоволително	65	54,0	11,0	121,0	2,24
еднократно × умерено	70	66,0	4,0	16,0	0,24
еднократно × пълно	15	30,0	−15,0	225,0	7,50
месечно × незадоволително	100	90,0	10,0	100,0	1,11
месечно × умерено	110	110,0	0,0	0,0	0,00
месечно × пълно	40	50,0	−10,0	100,0	2,00
2 седмици × незадоволително	15	36,0	−21,0	441,0	12,25
2 седмици × умерено	40	44,0	−4,0	16,0	0,36
2 седмици × пълно	45	20,0	25,0	625,0	31,25
Общо	**500**	**500**	**0,0**		**56,96**

$$\chi^2 = 2{,}24 + 0{,}24 + 7{,}50 + 1{,}11 + 0{,}00 + 2{,}00 + 12{,}25 + 0{,}36 + 31{,}25 = 56{,}96$$

Проверка

Сборът на колоната $O - E$ винаги е нула. Ако при вас не е, някоя очаквана честота е сметната погрешно.

Приносът на отделните клетки показва **къде** се намира връзката (Фиг. 5.3).



Фиг. 5.3: Принос на отделните клетки към стойността на хи-квадрат. Около три четвърти от общата стойност идват от двете клетки на най-интензивния режим: наблюдават се много повече пълни възстановявания и много по-малко незадоволителни резултати, отколкото се очаква при независимост.

5.4.5 Стъпка 5. Степени на свобода и p-стойност

$$df = (R - 1) \times (C - 1)$$

където R е броят категории по редове, а C — броят категории по колони. Степените на свобода **не зависят** от обема на извадката, а само от размера на таблицата.

В примера таблицата е 3 × 3:

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

Степените на свобода определят кой ред от таблицата за χ^2-разпределението се използва (Секция B).

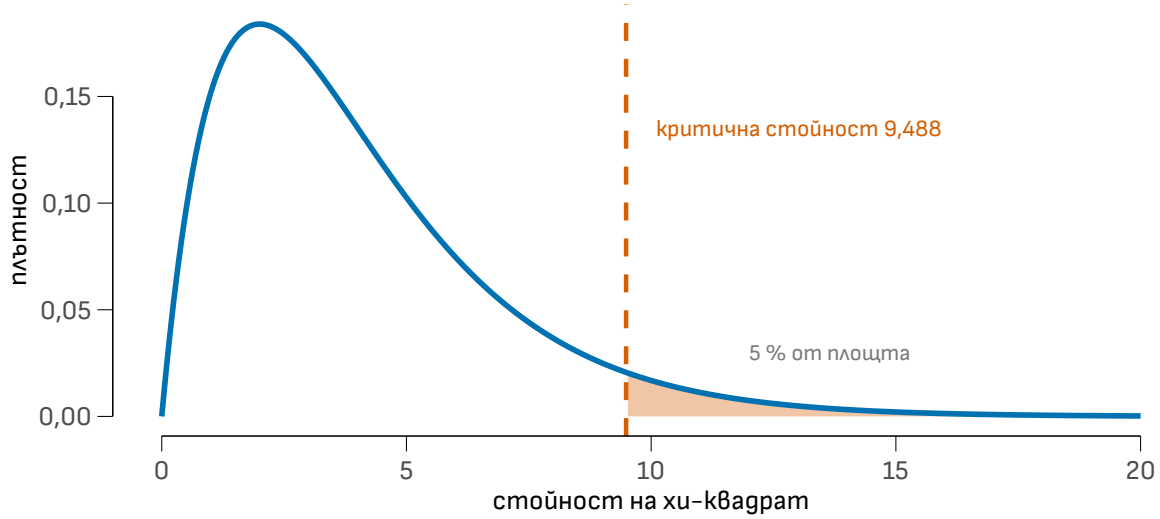
Таблица 5.5: Критични стойности на χ^2 при $df = 4$

df	$p = 0{,}10$	$p = 0{,}05$	$p = 0{,}01$
4	7,779	9,488	13,277

Изчисленото $\chi^2 = 56{,}96$ е по-голямо от 13,277 — най-дясната критична стойност. По таблицата записваме:

$$\chi^2 = 56{,}96 \text{ при } df = 4 \Rightarrow p < 0{,}01$$

Статистическият софтуер дава $p < 0{,}001$.



Фиг. 5.4: Разпределение на хи-квадрат при df = 4. Оцветената площ вдясно от критичната стойност 9,488 съответства на 5 % вероятност. Изчислената стойност 56,96 остава далеч извън обхвата на графиката, тоест p-стойността е много малка.

> **!** Посоката на теста
>
> Колкото по-голяма е изчислената χ^2-статистика, толкова по-малка е p-стойността и толкова по-силни са доказателствата срещу нулевата хипотеза. Отхвърляне се получава само при **големи** стойности, затова критичната област е винаги в десния край на разпределението.

5.4.6 Стъпка 6. Сила на връзката

χ^2-статистиката отговаря на въпроса *има ли връзка*, но не и *колко силна е тя*. Стойността ѝ расте пропорционално на обема на извадката: при достатъчно голямо N дори нищожна зависимост става статистически значима. Затова резултатът винаги се придружава от **коефициент на съчетаемост**.

$$V = \sqrt{\frac{\chi^2}{N \cdot (k - 1)}}$$

където N е общият брой наблюдения, а $k = \min(R,\ C)$ — по-малкото от броя редове и броя колони.

В примера $k = \min(3,\ 3) = 3$, следователно $k - 1 = 2$:

$$V = \sqrt{\frac{56{,}96}{500 \cdot 2}} = \sqrt{0{,}0570} = 0{,}24$$

V приема стойности от 0 (пълна независимост) до 1 (пълна зависимост). Ориентировъчните граници зависят от $k - 1$ (Таблица 5.6).

Таблица 5.6: Ориентировъчна интерпретация на V на Cramér

Сила на връзката	$k - 1 = 1$	$k - 1 = 2$	$k - 1 = 3$
слаба	0,10	0,07	0,06
умерена	0,30	0,21	0,17
силна	0,50	0,35	0,29

При $k - 1 = 2$ стойността 0,24 съответства на **умерена** връзка.

> **Пълен запис на резултата**
>
> Има статистически значима, умерена по сила връзка между честотата на физиотерапевтичното лечение и степента на възстановяване ($\chi^2 = 56{,}96$; $df = 4$; $p < 0{,}001$; $V = 0{,}24$; $N = 500$).
> Анализът на приносите показва, че връзката се дължи преди всичко на режима „2 седмици месечно", при който пълните възстановявания са повече, а незадоволителните резултати — по-малко от очакваните.

5.5 Решен пример: таблица 2 × 2 и коефициент φ

При таблица 2 × 2 се използва коефициентът φ. Той има същия смисъл като V, но носи и **знак**, който показва посоката на връзката, и се интерпретира като корелационен коефициент (Таблица 6.2).

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

където a, b, c и d са четирите наблюдавани честоти, разположени както в Таблица 5.7.

Таблица 5.7: Означения на клетките при таблица 2 × 2

	Изход „да"	Изход „не"
Фактор „да"	a	c
Фактор „не"	b	d

Условие

Родилно отделение регистрира данни за 200 едноплодни бременности. За всяка майка са известни тютюнопушенето по време на бременността и теглото на новороденото.

	Ниско тегло (< 2 500 г)	Нормално тегло	Всичко
Пушачка	a = 30	c = 50	80
Непушачка	b = 15	d = 105	120
Всичко	**45**	**155**	**200**

Задача. Има ли връзка между тютюнопушенето по време на бременността и ниското тегло при раждане и колко силна е тя?

Стъпка 1. Хипотези. H_0: тютюнопушенето и теглото при раждане са независими. H_1: не са независими. Ниво на значимост $\alpha = 0{,}05$.

Стъпка 2. Описание. Сред 80-те пушачки 30 са родили дете с ниско тегло ($37{,}5\ \%$), а сред 120-те непушачки — 15 ($12{,}5\ \%$).

Стъпка 3. Очаквани честоти.

$$E_{11} = \frac{80 \times 45}{200} = 18{,}0 \qquad E_{12} = \frac{80 \times 155}{200} = 62{,}0$$

$$E_{21} = \frac{120 \times 45}{200} = 27{,}0 \qquad E_{22} = \frac{120 \times 155}{200} = 93{,}0$$

Всички очаквани честоти са ≥ 5 и условията са изпълнени.

Стъпка 4. Тестова статистика.

$$\chi^2 = \frac{(30-18)^2}{18} + \frac{(50-62)^2}{62} + \frac{(15-27)^2}{27} + \frac{(105-93)^2}{93}$$

$$\chi^2 = 8{,}00 + 2{,}32 + 5{,}33 + 1{,}55 = 17{,}20$$

Стъпка 5. Степени на свобода и p-стойност.

$$df = (2-1) \times (2-1) = 1$$

df	$p = 0{,}10$	$p = 0{,}05$	$p = 0{,}01$
1	2,706	3,841	6,635

Понеже $17{,}20 > 6{,}635$, следва $p < 0{,}01$; софтуерът дава $p < 0{,}001$.

Стъпка 6. Сила и посока на връзката.

$$\varphi = \frac{(30 \cdot 105) - (15 \cdot 50)}{\sqrt{(30+15)(50+105)(30+50)(15+105)}}$$

$$\varphi = \frac{3\,150 - 750}{\sqrt{45 \cdot 155 \cdot 80 \cdot 120}} = \frac{2\,400}{\sqrt{66\,960\,000}} = \frac{2\,400}{8\,182{,}9} = 0{,}29$$

Заключение. Има статистически значима, умерена по сила връзка между тютюнопушенето по време на бременността и ниското тегло при раждане ($\chi^2 = 17{,}20$; $df = 1$; $p < 0{,}001$; $\varphi = 0{,}29$). Положителният знак показва, че тютюнопушенето се свързва с по-висока честота на ниско тегло при раждане.

i Връзката между φ, V и Z

При таблица 2 × 2 е изпълнено $k - 1 = 1$ и тогава

$$V = \sqrt{\frac{\chi^2}{N}} = |\varphi|$$

Проверка за примера: $\sqrt{17{,}20/200} = \sqrt{0{,}086} = 0{,}29$.
Освен това $Z = \sqrt{\chi^2} = \sqrt{17{,}20} = 4{,}15$ – същата стойност, която дава Z-тестът за двата дяла 37,5 % и 12,5 %.
Тоест V на Cramér е обобщението на φ за таблици с произволен размер.

5.6 Какво χ^2-тестът не прави

⚠ Четири ограничения

1. **Не доказва причинност.** Значимият резултат означава само, че двете променливи не са независими. Причинният извод изисква подходящ дизайн, а не по-малка p-стойност.
2. **Не посочва посока при таблици, по-големи от 2 × 2.** За да се разбере къде е връзката, се разглеждат приносите по клетки и процентите по редове.
3. **Не използва подредбата на категориите.** Ако едната променлива е ординална (стадий I–IV), тестът я третира като номинална и губи информация. За такива случаи съществуват тестове за тенденция.
4. **Не е приложим при свързани наблюдения.** При двукратно измерване на едни и същи лица се използва тестът на McNemar.

! Групировката влияе на извода

Заключението понякога зависи от начина, по който са групирани категориите. Обединяването на две съседни категории може да промени и стойността на χ^2, и извода. Затова групировката се определя от съдържателни съображения **преди** анализа, а не

след като се види резултатът.

5.7 Чести грешки

Таблица 5.8: Чести грешки при χ^2-теста

#	Грешка	Как да я избегнете
1	Изчисляване на χ^2 върху проценти	Тестът работи само с брой случаи
2	Забравяне на деленето на E	Всяка клетка се дели на **своята** очаквана честота
3	df, изчислено от обема на извадката	df зависи само от размера на таблицата
4	Прилагане при малки очаквани честоти	Проверете правилото за 80 % и за $E \geq 1$
5	Прилагане при свързани наблюдения	При двукратно измерване се използва тест на McNemar
6	Извод за причинност	χ^2 показва връзка, не причина
7	Съобщаване на χ^2 без сила на връзката	Винаги добавяйте V или φ
8	Извод за отделна категория при таблица 3×3	Първо погледнете приносите по клетки

5.8 Задачи за самостоятелна работа и домашна работа

Задача 5.1

Ретроспективно проучване анализира усложненията след аборт, извършен по два метода. Има ли зависимост между метода за аборт и вида на усложнението?

Метод за аборт	Кръвотечение	Инфекция	Инфертилитет	Общо
Класически	10	26	24	60
Аспирационен	12	13	15	40
Общо	**22**	**39**	**39**	**100**

Изчислете очакваните честоти, χ^2, df и V и формулирайте извод.

Задача 5.2 (домашна работа)

Има ли зависимост между продължителността на трудовия стаж при IT специалисти и наличието на неврологични оплаквания?

Продължителност на стажа	С оплаквания	Без оплаквания	Общо
До 5 г.	4	46	50
6–10 г.	15	85	100
11–15 г.	27	73	100
Над 15 г.	28	22	50
Общо	**74**	**226**	**300**

Освен теста посочете и коя категория допринася най-много за резултата.

Задача 5.3 (домашна работа)

Има ли зависимост между продължителността на трудовия стаж при IT специалисти и наличието на **зрителни** оплаквания?

Продължителност на стажа	Със зрителни оплаквания	Без оплаквания	Общо
До 5 г.	4	27	31
6–10 г.	6	46	52
11–15 г.	13	85	98
Над 15 г.	21	84	105
Общо	**44**	**242**	**286**

Сравнете извода с този от задача 5.2 и обяснете разликата.

Задача 5.4

В проучване на 250 пациенти с остър коронарен синдром е регистрирано дали реперфузионното лечение е започнало в първите 2 часа и дали пациентът е развил сърдечна недостатъчност.

	Сърдечна недостатъчност	Без недостатъчност	Общо
Реперфузия ≤ 2 ч.	18	132	150
Реперфузия > 2 ч.	32	68	100
Общо	**50**	**200**	**250**

Изчислете χ^2, df и φ и формулирайте извод, включително посоката на връзката.

Задача 5.5

В таблица 4 × 3 е получено $\chi^2 = 18{,}0$ при $N = 400$.
а) Колко са степените на свобода? б) Значим ли е резултатът при $\alpha = 0{,}05$? в) Изчислете V на Cramér и опишете силата на връзката.

Задача 5.6

Проучване с 60 участници дава таблица 2 × 2, в която най-малката **очаквана** честота е 3,2. Може ли да се приложи χ^2-тест? Какъв е подходящият избор?

Задача 5.7 (за разсъждение)

Проучване със 100 000 участници установява връзка между два признака с $\chi^2 = 12{,}0$, $df = 1$ и $V = 0{,}011$. Резултатът е статистически значим при $\alpha = 0{,}05$. Клинично значим ли е? Обяснете разликата между двете понятия в конкретния случай.

Задача 5.8 (за разсъждение)

Изследовател проверява последователно 20 независими връзки между качествени признаци при едно и също ниво $\alpha = 0{,}05$ и съобщава само двете, при които е получил $p < 0{,}05$. Защо този начин на представяне е подвеждащ?

5.9 Кратък тест

За всеки въпрос изберете **един** верен отговор. Ключът е в Секция Г.5.

1. χ^2-тестът за независимост се прилага, когато:

   а) и двете променливи са количествени;
   б) едната е количествена, а другата качествена;
   в) и двете променливи са качествени;
   г) променливата е една, но е измерена двукратно.

2. Очакваната честота за клетка се изчислява като:

   а) сума по ред плюс сума по колона, делено на N;
   б) произведение на сумата по ред и сумата по колона, делено на N;
   в) общият брой наблюдения, делен на броя клетки;
   г) наблюдаваната честота, умножена по N.

3. За таблица 4 × 3 степените на свобода са:

   а) 12;
   б) 7;
   в) 6;
   г) равни на $N - 1$.

4. Кое твърдение за χ^2-статистиката е вярно?

   а) Тя може да приема отрицателни стойности;
   б) Тя не зависи от обема на извадката;
   в) При равни други условия расте с увеличаване на извадката;
   г) Тя пряко измерва силата на връзката.

5. V на Cramér се съобщава, защото:

   а) заменя p-стойността;
   б) измерва силата на връзката независимо от обема на извадката;
   в) показва посоката на връзката при таблица 3 × 3;
   г) доказва причинна зависимост.

6. При таблица 2 × 2 и $N = 200$ е получено $\chi^2 = 8{,}0$. Стойността на V е приблизително:

   а) 0,04;
   б) 0,20;
   в) 0,40;
   г) не може да се изчисли.

6 Корелационен анализ

Цел на упражнението

След това упражнение трябва да умеете:

1. да разпознавате кога се прилага корелационен анализ и кога той е неподходящ;
2. да построявате и разчитате диаграма на разсейване;
3. да изчислявате коефициента на корелация на Pearson r стъпка по стъпка;
4. да описвате словесно силата и посоката на връзката;
5. да проверявате статистическата значимост на корелацията и да формулирате пълен извод;
6. да обяснявате защо корелацията не е доказателство за причинност.

Какво трябва да знаете отпреди

От Секция 2: средна аритметична, отклонение от средната, сума на квадратите на отклоненията. От Секция 4: нулева и алтернативна хипотеза, степени на свобода, p-стойност.

6.1 Кога какъв анализ

Изборът на метод за изследване на връзка се определя от вида на двете променливи (Таблица 6.1).

Таблица 6.1: Избор на метод за изследване на връзка

Първа променлива	Втора променлива	Метод
качествена	качествена	χ^2-тест, φ, V на Cramér (Секция 5)
количествена	качествена с 2 категории	t-тест (Секция 4)
количествена	количествена	**корелационен анализ** — сила и посока (тази глава)
количествена	количествена	**регресионен анализ** — модел за предсказване (Секция 7)

Определение

Корелационният анализ е статистически метод за изследване на взаимовръзката между две **количествени** променливи. Той оценява доколко двете променливи се променят съгласувано.

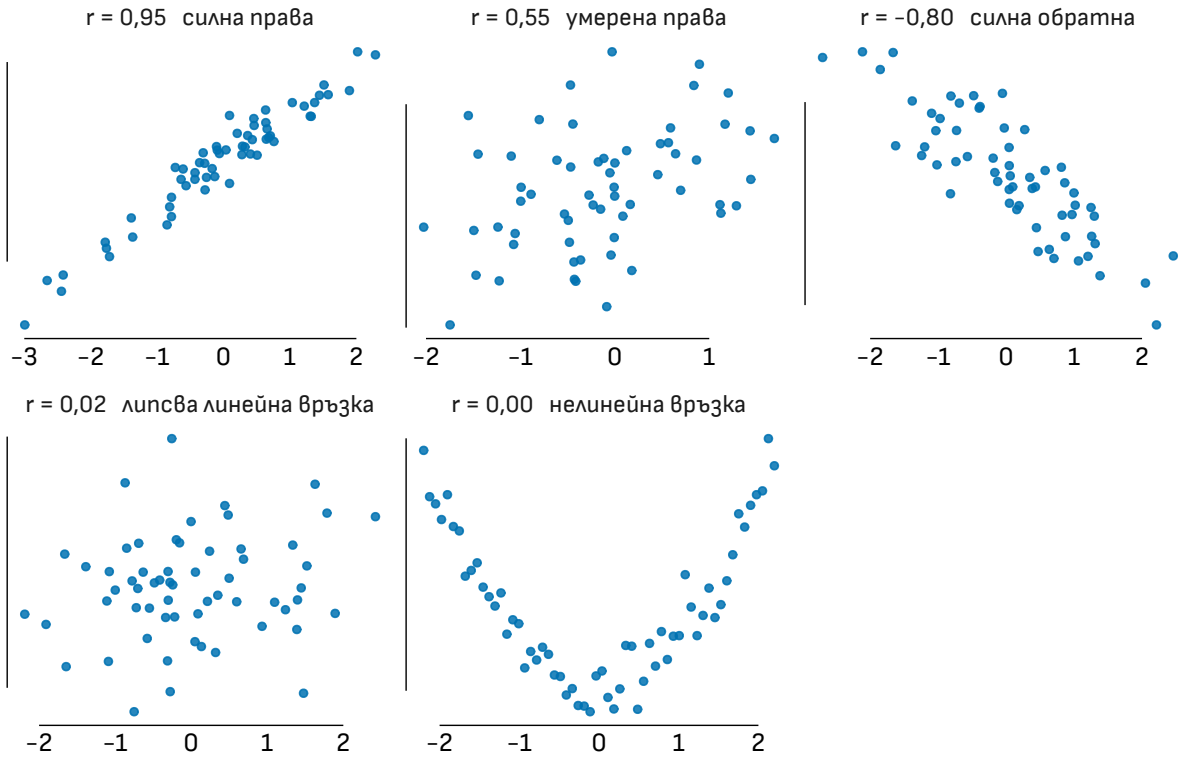
Корелация или регресия

Двата метода работят с едни и същи данни, но отговарят на различни въпроси.

- **Корелацията** отговаря на въпроса *колко силна и в каква посока е връзката*. Двете променливи са равноправни: коефициентът между X и Y е същият като между Y и X.
- **Регресията** отговаря на въпроса *как да предскажем едната променлива от другата*. Тук вече има зависима и независима променлива и размяната им променя резултата.

6.2 Първата стъпка е графиката

Корелационният анализ **започва** с диаграма на разсейване, а не с формулата. Всяко наблюдение се нанася като точка с координати $(X_i,\ Y_i)$. Графиката показва три неща, които числото само по себе си не показва: формата на връзката, наличието на екстремни стойности и еднородността на групата.



Фиг. 6.1: Диаграми на разсейване при различни стойности на r. В долния десен панел между двете променливи има ясна, но нелинейна връзка; коефициентът на Pearson я отчита като почти нулева. Затова графиката се разглежда преди изчислението.

6.3 Идеята зад коефициента

Нанесете средните аритметични $\bar{X}$ и $\bar{Y}$ като две линии върху диаграмата. Те разделят полето на четири квадранта (Фиг. 6.2).

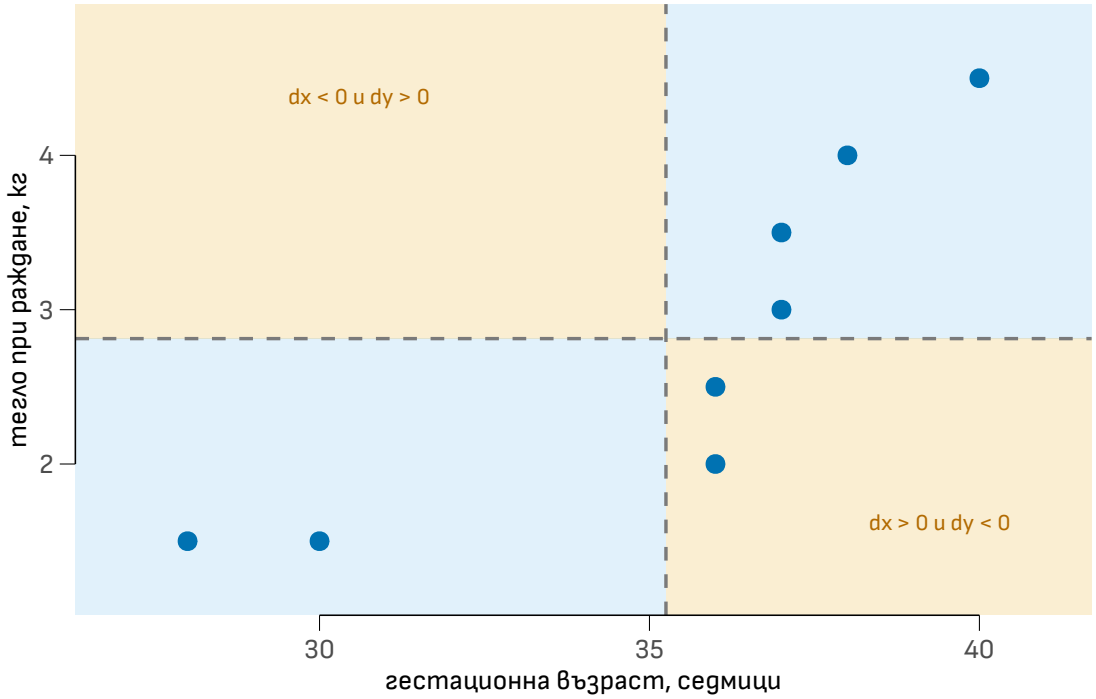
За всяко наблюдение се изчисляват двете отклонения:

$$d_x = X_i - \bar{X}, \qquad d_y = Y_i - \bar{Y}$$

и техният **кръстосан продукт** $d_x \cdot d_y$.

- В горния десен квадрант и двете отклонения са положителни, така че произведението е **положително**.
- В долния ляв квадрант и двете отклонения са отрицателни, така че произведението отново е **положително**.
- В другите два квадранта знаците са различни и произведението е **отрицателно**.

Сумата $\sum(d_x \cdot d_y)$ е положителна, когато точките се струпват по диагонала от долу вляво към горе вдясно (права връзка), и отрицателна при обратната подредба.



Фиг. 6.2: Четирите квадранта около средните аритметични. Точките в сините квадранти дават положителен принос към сумата на кръстосаните продукти, а точките в оранжевите — отрицателен. При права връзка преобладават сините.

6.4 Коефициент на Pearson

Ако използвахме само сумата $\sum(d_x \cdot d_y)$, стойността ѝ щеше да зависи от мерните единици: същите данни, изразени в грамове вместо в килограми, биха дали хиляда пъти по-голямо число. Затова сумата се дели на израз, който отстранява мащаба:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \cdot \sum(Y_i - \bar{Y})^2}} = \frac{\sum(d_x \cdot d_y)}{\sqrt{\sum d_x^2 \cdot \sum d_y^2}}$$

където:

- X_i и Y_i са стойностите на двете променливи за i-тото наблюдение;
- $\bar{X}$ и $\bar{Y}$ — средните аритметични на двете променливи;
- $d_x \cdot d_y$ — кръстосаният продукт за всяко наблюдение;
- $\sum(d_x \cdot d_y)$ — сумата на кръстосаните продукти, тоест числителят;
- $\sum d_x^2$ и $\sum d_y^2$ — сумите на квадратите на отклоненията.

! Свойства на коефициента

1. r е **безразмерно** число и не зависи от мерните единици.
2. r приема стойности само в интервала от -1 до $+1$.
3. Знакът показва **посоката**, а абсолютната стойност — **силата** на връзката.
4. r между X и Y е равен на r между Y и X.
5. r измерва само **линейна** връзка.

Посока на връзката

Права (положителна) корелация: двете променливи се изменят в една и съща посока. Пример: гестационна възраст и тегло при раждане.
Обратна (отрицателна) корелация: увеличението на едната се свързва с намаление на другата. Пример: изминато разстояние при 6-минутен тест с ходене и ниво на NT-proBNP.

Таблица 6.2: Интерпретация на коефициента на Pearson

r	Сила и посока на връзката
$+1$	пълна (функционална) права
от $+0{,}67$ до $+1$	силна права
от $+0{,}34$ до $+0{,}66$	умерена права
от 0 до $+0{,}33$	слаба права
0	**липсва линейна връзка**
от 0 до $-0{,}33$	слаба обратна
от $-0{,}34$ до $-0{,}66$	умерена обратна
от $-0{,}67$ до -1	силна обратна
-1	пълна (функционална) обратна

! Две уточнения за $r = 0$ и $r = \pm 1$

Пълна (функционална) зависимост означава, че всяко определено изменение в едната променлива е свързано със строго определено изменение в другата, тоест може да се изведе точна формула $Y = f(X)$. В медицината такива връзки практически не се срещат.
$r = 0$ **не** означава, че между двете променливи няма никаква връзка. То означава, че ако съществува връзка, тя **не е линейна**. Точно затова диаграмата на разсейване се разглежда преди изчисляването на r.

⚠ Условия за прилагане на коефициента на Pearson

1. И двете променливи са количествени, измерени на интервална или пропорционална скала.
2. Наблюденията са **независими** — всяка двойка стойности принадлежи на различно лице.
3. Връзката е приблизително **линейна**, което се проверява графично.
4. Няма единични екстремни стойности, които сами определят резултата.
5. За проверката на значимост е желателно двете променливи да са приблизително нормално разпределени.

Ако условие 3 или 5 не е изпълнено, се използва рангов коефициент, например този на Spearman.

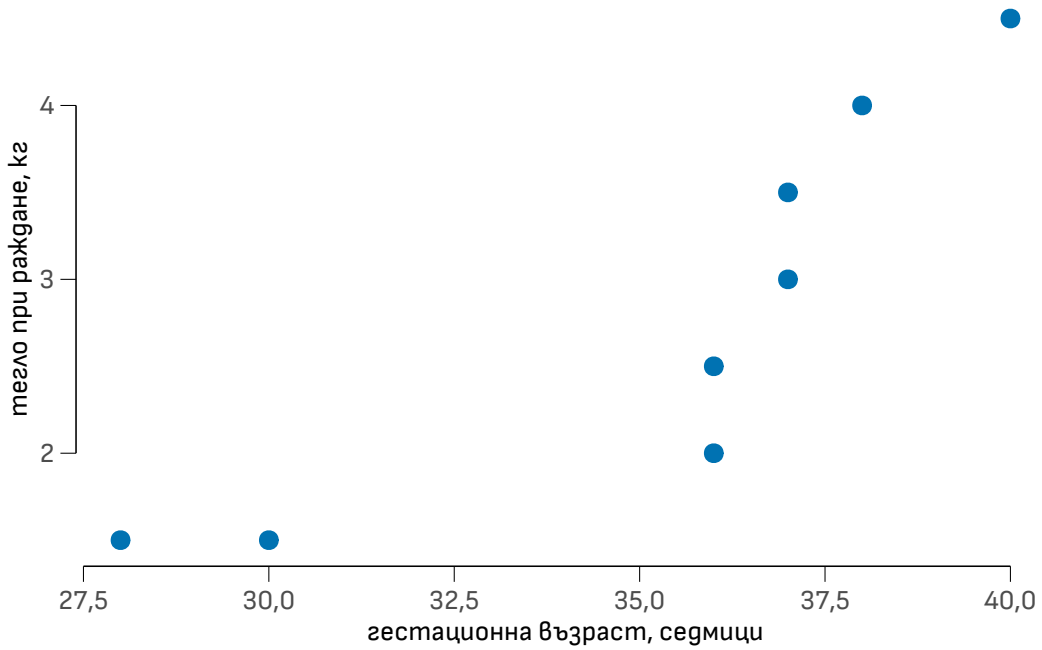
6.5 Решен пример

Условие

Регистрирани са гестационната възраст при раждане X (седмици) и теглото при раждане Y (kg) при 8 новородени.

Новородено №	Гестационна възраст X, седмици	Тегло Y, kg
1	28	1,5
2	30	1,5
3	36	2,0
4	36	2,5
5	37	3,0
6	37	3,5
7	38	4,0
8	40	4,5

Задача. Да се оцени силата и посоката на връзката между гестационната възраст и теглото при раждане.



Фиг. 6.3: Диаграма на разсейване за осемте новородени. Точките се струпват около една възходяща линия и не се откроява отделно наблюдение, което да определя само по себе си картината. Условията за коефициента на Pearson са изпълнени.

6.5.1 Стъпка 1. Средни аритметични

$$\bar{X} = \frac{\sum X}{n} = \frac{282}{8} = 35{,}25 \text{ седмици}$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{22{,}5}{8} = 2{,}8125 \text{ кг}$$

> **Не закръгляйте средните**
>
> Двете стойности са точни, а не приблизителни. Ако $\bar{X}$ се закръгли до 35, а $\bar{Y}$ до 3, коефициентът става 0,85 вместо 0,87 и вече не съвпада със стойността, която ще получим от регресионния модел в Секция 7. Работете с точните стойности и закръгляйте едва крайния резултат.

6.5.2 Стъпка 2. Отклонения, кръстосани продукти и квадрати

$$d_x = X - \bar{X}, \qquad d_y = Y - \bar{Y}$$

Таблица 6.3: Работна таблица за коефициента на корелация

№	X	Y	d_x	d_y	$d_x \cdot d_y$	d_x^2	d_y^2
1	28	1,5	−7,25	−1,3125	9,5156	52,5625	1,7227

№	X	Y	d_x	d_y	$d_x \cdot d_y$	d_x^2	d_y^2
2	30	1,5	−5,25	−1,3125	6,8906	27,5625	1,7227
3	36	2,0	0,75	−0,8125	−0,6094	0,5625	0,6602
4	36	2,5	0,75	−0,3125	−0,2344	0,5625	0,0977
5	37	3,0	1,75	0,1875	0,3281	3,0625	0,0352
6	37	3,5	1,75	0,6875	1,2031	3,0625	0,4727
7	38	4,0	2,75	1,1875	3,2656	7,5625	1,4102
8	40	4,5	4,75	1,6875	8,0156	22,5625	2,8477
Общо	**282**	**22,5**	**0,00**	**0,00**	**28,3750**	**117,5000**	**8,9688**

Две проверки

- Сборът на d_x и сборът на d_y винаги са нула. Ако при вас се получава друго число, някоя средна е сметната погрешно.
- Сумите $\sum d_x^2$ и $\sum d_y^2$ винаги са положителни. Отрицателна стойност е сигурен признак за грешка.

6.5.3 Стъпка 3. Изчисляване на r

$$r = \frac{\sum (d_x \cdot d_y)}{\sqrt{\sum d_x^2 \cdot \sum d_y^2}} = \frac{28{,}3750}{\sqrt{117{,}5000 \cdot 8{,}9688}}$$

$$r = \frac{28{,}3750}{\sqrt{1\,053{,}83}} = \frac{28{,}3750}{32{,}4627} = 0{,}87$$

6.5.4 Стъпка 4. Проверка на статистическата значимост

Изчисленият коефициент описва **само тези 8 новородени**. За да се провери дали връзката съществува и в генералната съвкупност, се прави тест на хипотеза.

H_0: Популационният корелационен коефициент е нула ($\rho = 0$) — между гестационната възраст и теглото при раждане няма линейна връзка.

H_1: Популационният корелационен коефициент е различен от нула ($\rho \neq 0$).

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}, \qquad df = n - 2$$

$$t = \frac{0{,}87 \cdot \sqrt{8-2}}{\sqrt{1-0{,}87^2}} = \frac{0{,}87 \cdot 2{,}4495}{\sqrt{1-0{,}7640}} = \frac{2{,}1411}{0{,}4858} = 4{,}41$$

$$df = 8 - 2 = 6$$

Таблица 6.4: Критични стойности на t при $df = 6$

df	$p = 0{,}10$	$p = 0{,}05$	$p = 0{,}01$
6	1,943	2,447	3,707

Изчисленото $|t| = 4{,}41$ е по-голямо от 3,707, следователно $p < 0{,}01$; софтуерът дава $p = 0{,}005$.

> **Защо малката извадка изисква висок коефициент**
>
> При $n = 8$ и следователно $df = 6$ значимост се постига едва при $|r| \approx 0{,}71$. При $n = 30$ същият праг е около 0,36, а при $n = 100$ — около 0,20. Тоест **едно и също** число r може да бъде значимо при голяма извадка и незначимо при малка. По тази причина се съобщават и r, и n, и p.

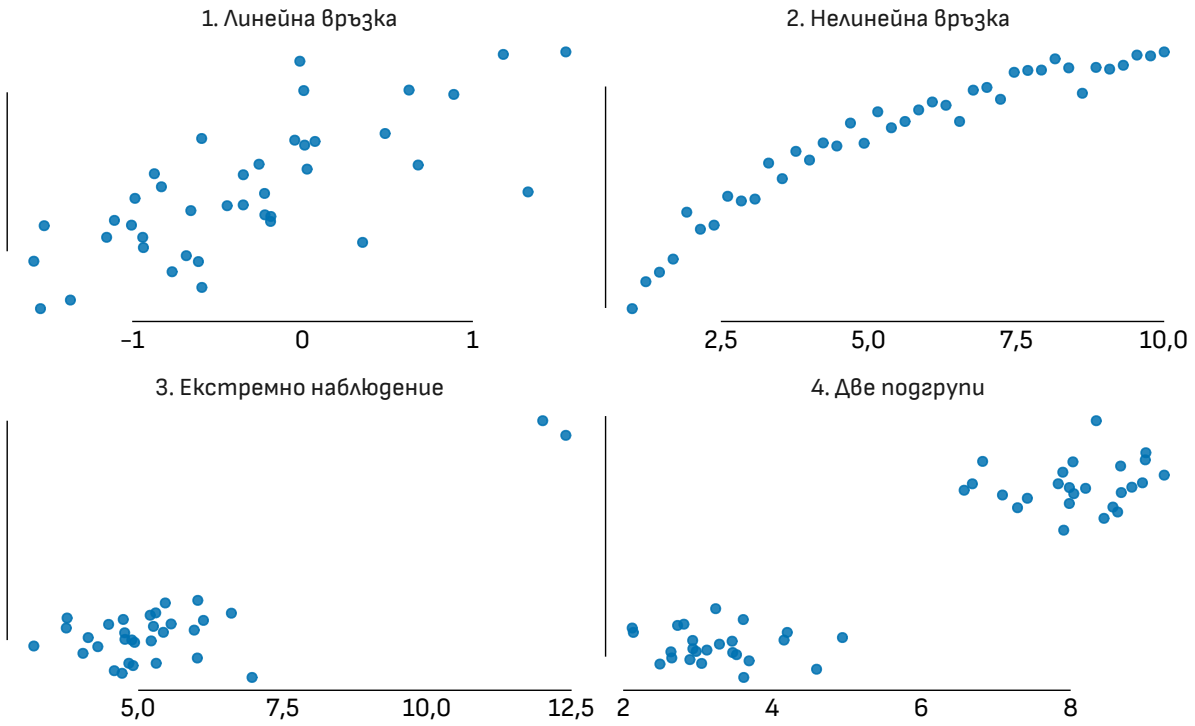
6.5.5 Стъпка 5. Заключение

$r = 0{,}87$. Съществува **силна права** връзка между гестационната възраст и теглото при раждане при тези 8 новородени: с увеличаване на гестационната възраст се увеличава и теглото при раждане. Връзката е статистически значима ($t = 4{,}41$; $df = 6$; $p < 0{,}01$).

6.6 Какво коефициентът не показва

Едно число не може да замени графиката. Четирите панела на Фиг. 6.4 имат почти еднакви коефициенти на корелация, но данните зад тях са съвсем различни.

> **Пет капана**
>
> 1. **Нелинейност.** Коефициентът измерва само линейната част от връзката.
> 2. **Екстремни наблюдения.** Едно-единствено далечно наблюдение може да създаде или да разруши висок коефициент.
> 3. **Смесени подгрупи.** Ако извадката съдържа две различни популации, коефициентът описва разстоянието между тях, а не връзка вътре в тях.
> 4. **Стеснен обхват.** Ако измерваме гестационната възраст само между 39 и 41 седмица, коефициентът ще бъде близък до нула, макар в целия диапазон връзката да е силна.
> 5. **Екологична заблуда.** Корелация, установена между **средни стойности за държави или региони**, не важи автоматично за отделните лица.



Фиг. 6.4: Четири набора данни с почти еднакъв коефициент на корелация около 0,8. Само в първия панел числото описва вярно данните: в останалите то е резултат съответно на нелинейност, на едно екстремно наблюдение и на две разделени подгрупи.

6.6.1 Екстраполация

> ⚠ Определение
>
> **Екстраполацията** е разпростиране на установената връзка върху стойности **извън** обхвата на данните, върху които тя е построена.

В изваДка предимно от деца се наблюдава силна връзка между теглото и ръста. Тази връзка обаче не бива да се пренася върху възрастните: при тях растежът на височина е завършил, а теглото може да продължи да се увеличава. Връзката е валидна само в обхвата на наличните данни.

6.7 Корелация и причинност

> ❗ Най-важното изречение в тази глава
>
> Доказването на статистическа зависимост **не** означава доказване на причинно-следствена връзка.

Наличието на корелация между X и Y допуска поне пет обяснения:

1. X причинява Y;

2. Y причинява X (обратна причинност);
3. трета променлива Z причинява и двете (замъгляващ фактор, вж. Секция 1.3);
4. връзката се дължи на начина, по който е подбрана извадката;
5. връзката е случайна.

Статистиката сама по себе си не може да избере между тези обяснения. Вероятността връзката да е причинно-следствена е по-голяма, когато:

1. връзката е **силна**;
2. връзката е **биологично правдоподобна** — има известен механизъм;
3. връзката се потвърждава и в **други, независими проучвания**;
4. **времевата последователност** е правилна: причината предхожда следствието;
5. се наблюдава **зависимост от дозата** — по-голяма експозиция води до по-голям ефект.

> **Липсата на механизъм не отхвърля причинността**
>
> Ако механизмът не е известен, това означава само, че познанието е непълно. Връзката между тютюнопушенето и белодробния карцином е установена статистически десетилетия преди механизмът да бъде изяснен. Обратно, съществуването на правдоподобен механизъм само по себе си не доказва причинност.

6.8 Чести грешки

Таблица 6.5: Чести грешки при корелационния анализ

#	Грешка	Как да я избегнете
1	Корелация между качествени признаци	Pearson изисква две количествени променливи
2	Извод „няма връзка“ при $r = 0$	Означава липса само на **линейна** връзка
3	Изчисляване на r без поглед към графиката	Диаграмата на разсейване се строи първа
4	Твърдение за причинност	Употребявайте „свързано с“, а не „причинява“
5	Пропускане на проверката за значимост	r описва извадката; изводът за популацията изисква тест
6	Грубо закръгляване на средните	Води до различен r и до противоречие с R^2
7	Съобщаване на r без n и p	Едно и също r означава различно при различен обем
8	Екстраполация извън обхвата на данните	Връзката важи само в наблюдавания диапазон

6.9 Задачи за самостоятелна работа и домашна работа

Задача 6.1

Да се измери силата и посоката на зависимостта между възрастта при поставяне на диагноза и преживяемостта при пациенти с белодробен карцином.

Пациент №	1	2	3	4	5	6	7	8
Възраст X, години	71	71	73	74	75	76	76	78
Преживяемост Y, месеци	135	96	66	44	16	22	11	5

Проверете и статистическата значимост на получения коефициент.

Задача 6.2 (домашна работа)

В таблицата са отчетени изминатото разстояние при 6-минутен тест с ходене (6MWT) и нивото на NT-proBNP при 8 пациенти с начална сърдечна недостатъчност. Да се измери силата и посоката на зависимостта.

Пациент №	1	2	3	4	5	6	7	8
Разстояние X, км	0,7	0,7	0,6	0,5	0,4	0,4	0,3	0,3
NT-proBNP Y, pg/µl	110	120	140	160	180	200	240	300

Задача 6.3 (домашна работа)

Постройте диаграма на разсейване за данните от задача 6.2 и посочете:
а) каква е формата на връзката; б) има ли наблюдение, което се отклонява от общата тенденция; в) подходящ ли е коефициентът на Pearson за тези данни.

Задача 6.4

Възможно ли е при дадена извадка всички индивидуални отклонения d_x да бъдат положителни? А всички да бъдат равни на нула? Какъв би бил коефициентът на корелация във втория случай и защо?

Задача 6.5

Проучване съобщава $r = -0{,}45$ между продължителността на съня и нивото на кортизол при 60 участници.
а) Опишете с думи силата и посоката на връзката. б) Проверете значимостта при $\alpha = 0{,}05$. в) Как ще се промени изводът, ако същият коефициент е получен при 10 участници?

Задача 6.6 (за разсъждение)

Проучване установява $r = 0{,}62$ между броя изядени сладоледи на ден и броя на удавянията през летните месеци. Означава ли това, че сладоледът причинява удавяния? Кое от петте обяснения за корелация е най-вероятното тук?

Задача 6.7 (за разсъждение)

Изследовател измерва връзката между ръст и тегло само сред професионални баскетболисти и получава $r = 0{,}12$. Той заключава, че при възрастни ръстът и теглото не са свързани. Кой от петте капана е допуснат?

Задача 6.8

Две проучвания съобщават една и съща стойност $r = 0{,}30$. Първото е с $n = 20$, второто — с $n = 200$. Изчислете тестовата статистика и за двете и обяснете защо изводите се различават.

6.10 Кратък тест

За всеки въпрос изберете **един** верен отговор. Ключът е в Секция Г.6.

1. Коефициентът на Pearson се прилага, когато:

   а) и двете променливи са качествени;
   б) и двете променливи са количествени;
   в) едната е качествена, а другата количествена;
   г) променливата е една, измерена двукратно.

2. Стойност $r = 0$ означава, че:

   а) между двете променливи няма никаква връзка;
   б) между двете променливи няма линейна връзка;
   в) двете променливи са независими;
   г) извадката е твърде малка.

3. Коефициентът $r = -0{,}82$ описва връзка, която е:

   а) слаба права;
   б) умерена обратна;
   в) силна обратна;
   г) пълна обратна.

4. Кое **не** променя стойността на коефициента на корелация?

   а) Преминаване от килограми към грамове;
   б) Добавяне на едно екстремно наблюдение;
   в) Стесняване на обхвата на едната променлива;
   г) Обединяване на две различни подгрупи.

5. При $n = 8$ и $r = 0{,}87$ степените на свобода за проверката на значимост са:

   а) 8;
   б) 7;
   в) 6;
   г) 2.

6. Статистически значима корелация между два признака означава, че:

   а) единият признак причинява другия;
   б) връзката най-вероятно не се дължи само на случайност;
   в) връзката е клинично важна;
   г) връзката е задължително линейна и силна.

7 Регресионен анализ

Цел на упражнението

След това упражнение трябва да умеете:

1. да различавате зависима от независима променлива и да обосновавате избора;
2. да обяснявате какво минимизира правата на най-добро съответствие;
3. да изчислявате регресионния коефициент b и свободния член a;
4. да записвате регресионното уравнение и да го използвате за прогнозиране;
5. да интерпретирате коефициента на детерминация R^2 и да извеждате от него коефициента на корелация заедно с неговия знак;
6. да посочвате границите, в които моделът е валиден.

Какво трябва да знаете отпреди

От Секция 6: диаграма на разсейване, отклонения от средната, сила и посока на връзката, коефициент на Pearson.

7.1 Същност на регресионния модел

Корелационният анализ отговаря на въпроса *колко силна е връзката*. Регресионният анализ отива една крачка по-нататък и отговаря на въпроса *как да предскажем едната променлива от другата*. За целта връзката се свежда до уравнение.

Определения

Зависима (резултативна) променлива Y — тази, чиито стойности се предсказват.
Независима (факторна) променлива X — тази, която се използва за предсказването; нарича се още предиктор или фактор.

Уравнението на **простата (единичната) линейна регресия** е:

$$Y = b \cdot X + a + \varepsilon$$

където:

- b е **регресионният коефициент** — средната промяна в Y при увеличаване на X с **една единица**;
- a е **свободният член** (константа, отсечка по оста Y) — стойността на Y при $X = 0$;
- ε е **остатъчният компонент** — частта от Y, която моделът не обяснява.

Прогнозираната стойност се означава с $\hat{Y}$ (или Y_t) и се получава без остатъчния компонент:

$$\hat{Y} = b \cdot X + a$$

Как да разпознаете коя променлива е зависима

Питайте се какво е известно предварително и какво искаме да предскажем. Гестационната възраст е известна преди раждането, а теглото се измерва след него, следователно теглото е зависимата променлива. Обратната постановка е математически възможна, но няма практически смисъл.
За разлика от корелацията, при регресията размяната на X и Y **променя** резултата: получава се друга права.

Видове регресионни модели

Според формата на зависимостта: линейни, при които регресията е права линия, и нелинейни, при които тя е крива. В курса се разглеждат само линейните.
Според броя на факторите: еднофакторни (прости) с един предиктор и многофакторни (множествени) с два и повече предиктора, всеки със собствен регресионен коефициент.

7.2 Правата на най-добро съответствие

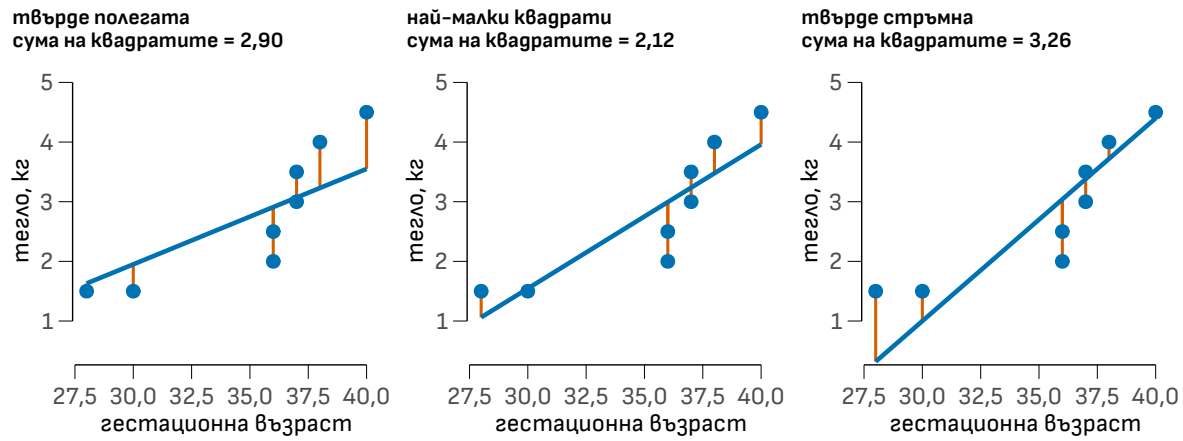
През едно облако от точки могат да се прекарат безброй прави. Регресионният анализ избира **една** от тях по ясен критерий.

За всяко наблюдение разстоянието по вертикала между действителната стойност Y_i и стойността върху правата $\hat{Y}_i$ се нарича **остатък**:

$$e_i = Y_i - \hat{Y}_i$$

Критерий на най-малките квадрати

Правата на най-добро съответствие е тази, за която **сумата от квадратите на остатъците** $\sum e_i^2$ е възможно най-малка.
Квадратирането служи за същото, за което и при стандартното отклонение: без него положителните и отрицателните остатъци биха се занулили взаимно и всяка права, минаваща през центъра на данните, би изглеждала еднакво добра.



Фиг. 7.1: Три възможни прави през едни и същи данни. За всяка е показана сумата от квадратите на остатъците. Методът на най-малките квадрати избира средната права, при която тази сума е най-малка.

7.3 Решен пример

Условие

Регистрирани са гестационната възраст при раждане X (седмици) и теглото при раждане Y (кг) при 8 новородени — същите данни, с които работихме в Секция 6. **Задача.** Да се моделира зависимостта на теглото при раждане от гестационната възраст и да се оцени качеството на модела.

7.3.1 Стъпка 1. Средни аритметични

$$\bar{X} = \frac{282}{8} = 35{,}25 \text{ седмици}, \qquad \bar{Y} = \frac{22{,}5}{8} = 2{,}8125 \text{ кг}$$

Точност на междинните изчисления

При регресията **не се препоръчва** закръгляне на междинните резултати. Закръгляването на $\bar{X}$ до 35 или на b до един знак променя забележимо свободния член и прогнозите. Закръглявайте едва крайния отговор.

7.3.2 Стъпка 2. Помощна таблица

Таблица 7.1: Помощна таблица за регресионния анализ

№	X	Y	$X \cdot Y$	X^2
1	28	1,5	42,0	784
2	30	1,5	45,0	900

№	X	Y	$X \cdot Y$	X^2
3	36	2,0	72,0	1 296
4	36	2,5	90,0	1 296
5	37	3,0	111,0	1 369
6	37	3,5	129,5	1 369
7	38	4,0	152,0	1 444
8	40	4,5	180,0	1 600
$n = 8$	**282**	**22,5**	**821,5**	**10 058**

7.3.3 Стъпка 3. Регресионен коефициент b

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

Числител:

$$8 \cdot 821{,}5 - 282 \cdot 22{,}5 = 6\,572 - 6\,345 = 227$$

Знаменател:

$$8 \cdot 10\,058 - 282^2 = 80\,464 - 79\,524 = 940$$

$$b = \frac{227}{940} = 0{,}2415 \approx 0{,}24 \text{ кг на седмица}$$

Положителната стойност на b означава **нарастващ наклон** на правата, тоест права (положителна) връзка.

> ⚠ Внимание в знаменателя
>
> $\sum X^2$ и $(\sum X)^2$ са **различни** величини. Първата е сборът на квадратите (10 058), втората — квадратът на сбора ($282^2 = 79\,524$). Разместването им е най-честата грешка в тази стъпка.

7.3.4 Стъпка 4. Свободен член a

$$a = \bar{Y} - (b \cdot \bar{X})$$

$$a = 2{,}8125 - (0{,}2415 \cdot 35{,}25) = 2{,}8125 - 8{,}5125 = -5{,}70$$

Отрицателен свободен член — проблем ли е?

Не. Формално $a = -5{,}70$ означава „тегло при гестационна възраст нула седмици", което е безсмислено. Свободният член няма самостоятелна клинична интерпретация, когато стойността $X = 0$ е извън обхвата на данните. Той определя откъде тръгва правата и е необходим за прогнозите в наблюдавания диапазон.

Свойство, което си струва да се знае

Правата на най-добро съответствие винаги минава през точката $(\bar{X},\ \bar{Y})$. Проверка за примера: $0{,}2415 \cdot 35{,}25 - 5{,}70 = 2{,}8125$ — точно средната стойност на Y.

7.3.5 Стъпка 5. Регресионно уравнение

$$\hat{Y} = b \cdot X + a$$

$$\boxed{\hat{Y} = 0{,}2415 \cdot X - 5{,}70}$$

За построяване на правата са достатъчни две точки:

Таблица 7.2: Две точки за построяване на правата

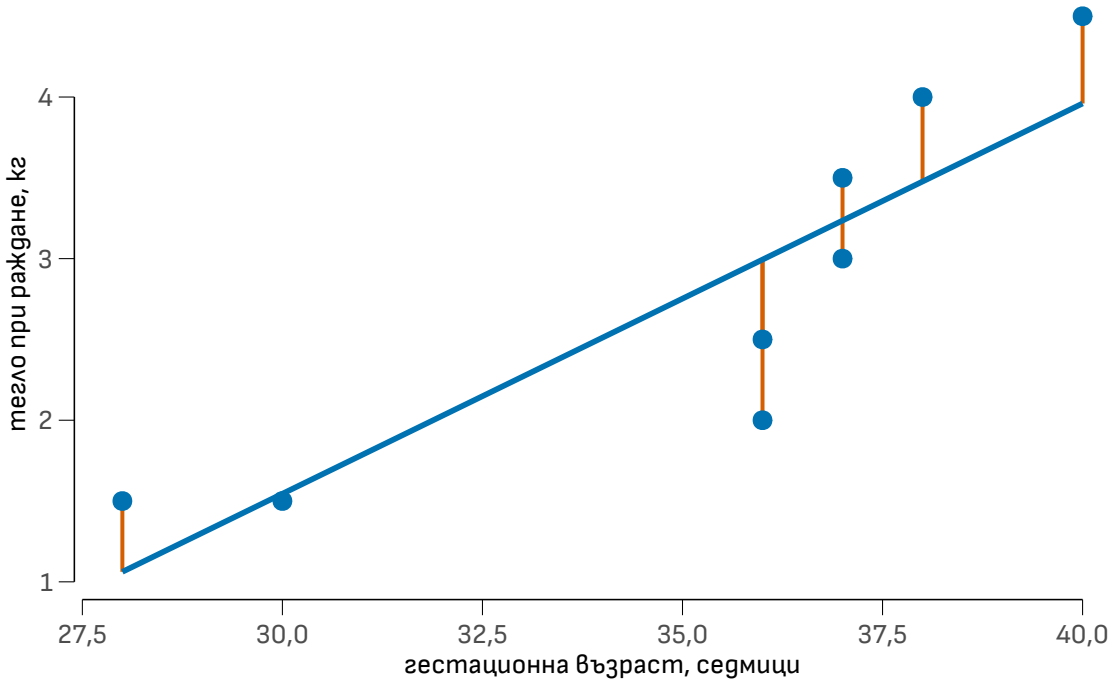
Гестационна възраст X	Прогнозирано тегло $\hat{Y} = 0{,}2415X - 5{,}70$
28 седмици	$0{,}2415 \cdot 28 - 5{,}70 = 1{,}06$ кг
40 седмици	$0{,}2415 \cdot 40 - 5{,}70 = 3{,}96$ кг

7.3.6 Стъпка 6. Остатъци

Остатъците показват доколко моделът се справя с всяко отделно наблюдение.

Таблица 7.3: Прогнозирани стойности и остатъци

№	X	Y	$\hat{Y}$	Остатък $e = Y - \hat{Y}$
1	28	1,5	1,06	0,44
2	30	1,5	1,54	−0,04
3	36	2,0	2,99	−0,99
4	36	2,5	2,99	−0,49
5	37	3,0	3,24	−0,24
6	37	3,5	3,24	0,26
7	38	4,0	3,48	0,52
8	40	4,5	3,96	0,54
Общо		**22,5**	**22,5**	**0,00**



Фиг. 7.2: Правата на най-добро съответствие. Червените отсечки са остатъците — разликите между действителното и прогнозираното тегло. Сборът на квадратите им е по-малък от този при всяка друга права.

> **Проверка**
>
> Сборът на остатъците при правилно изчислен модел е нула, а сборът на прогнозираните стойности е равен на сбора на наблюдаваните. Това е удобна контролна точка.

7.3.7 Стъпка 7. Коефициент на детерминация R^2

Всеки регресионен модел се оценява по способността му да предсказва правилно. Тази оценка се извършва чрез **коефициента на детерминация** R^2 (коефициент на определеност).

Идеята е разлагане на общото разсейване на зависимата променлива:

$$\underbrace{\sum (Y_i - \bar{Y})^2}_{\text{общо разсейване}} = \underbrace{\sum (\hat{Y}_i - \bar{Y})^2}_{\text{обяснено от модела}} + \underbrace{\sum (Y_i - \hat{Y}_i)^2}_{\text{необяснено (остатъчно)}}$$

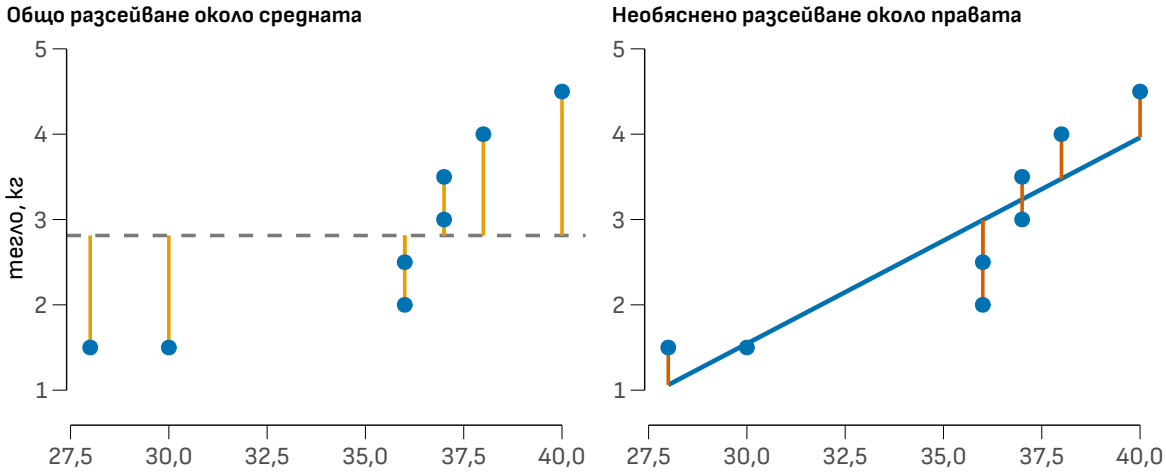
$$R^2 = \frac{\text{обяснено разсейване}}{\text{общо разсейване}}$$

За примера общото разсейване е $\sum d_y^2 = 8{,}9688$ (изчислено в Секция 6), а необясненото — $\sum e_i^2 = 2{,}1165$. Следователно:

$$R^2 = \frac{8{,}9688 - 2{,}1165}{8{,}9688} = \frac{6{,}8523}{8{,}9688} = 0{,}764$$

! Свойства на R^2

1. Приема стойности между 0 и 1 (или между 0 % и 100 %).
2. Показва **каква част** от разсейването на зависимата променлива се обяснява с действието на факторната.
3. Колкото е по-близо до 1, толкова по-точно моделът предсказва.
4. При проста линейна регресия е изпълнено $R^2 = r^2$ и следователно $|r| = \sqrt{R^2}$.



Фиг. 7.3: Разлагане на разсейването. Вляво отклоненията на наблюденията от общата средна (общо разсейване), вдясно отклоненията от регресионната права (необяснено разсейване). Колкото по-къси са отсечките вдясно спрямо тези вляво, толкова по-голямо е R2.

7.3.8 Стъпка 8. Заключение

- Всяка една седмица увеличение на гестационната възраст се асоциира с **0,24 кг** увеличение на очакваното тегло при раждане.
- **76,4 %** от вариабилността на теглото при раждане може да се обясни с гестационната възраст. Останалите 23,6 % се дължат на други фактори и на случайна вариация.
- Коефициентът на корелация е $r = +\sqrt{0{,}764} = +0{,}87$ – силна права връзка, което съвпада с резултата от Секция 6.

! Знакът на r идва от b

Коренуването на R^2 дава само **абсолютната стойност** $|r|$. Знакът се определя от знака на регресионния коефициент b, тоест от наклона на правата:

- $b > 0$ – правата се качва – r е **положителен**;
- $b < 0$ – правата се спуска – r е **отрицателен**.

Затова при въпрос „каква е стойността на r“ първо се поглежда наклонът.

7.4 Прогнозиране

Моделът се използва, като в уравнението се замести известната стойност на X.

Пример

Постъпва родилка в 34-та гестационна седмица. Какво тегло се очаква?

$$\hat{Y} = 0{,}2415 \cdot 34 - 5{,}70 = 8{,}211 - 5{,}70 = 2{,}51 \text{ кг}$$

Три ограничения на прогнозата

1. **Обхват на данните.** Моделът е построен върху наблюдения от 28 до 40 гестационна седмица. Прогноза за 20-та седмица е екстраполация: формулата ще даде число ($0{,}2415 \cdot 20 - 5{,}70 = -0{,}87$ кг), но то е безсмислено.
2. **Индивид срещу група.** $\hat{Y}$ е **средно очакваната** стойност за лица с даден X, а не точното тегло на конкретното новородено. Отделните стойности се разпръскват около правата.
3. **Причинност.** Регресията описва връзка. Твърдението, че промяна в X ще доведе до промяна в Y, изисква подходящ дизайн, а не по-висок R^2.

Регресионните модели стоят в основата на много клинични инструменти. Скалите за оценка на сърдечно-съдов риск и на риска от инсулт при предсърдно мъждене са изведени именно от регресионни модели, при които няколко фактора се комбинират в една прогноза.

7.5 Общо правило за интерпретация

1. **Регресионният коефициент** b показва как се променя резултатът при увеличаване на фактора с **една единица** — с нейното име: седмица, милиграм, минута.
2. **Положителен** b означава права връзка, **отрицателен** — обратна.
3. R^2 показва какъв дял от вариацията на резултативната променлива се обяснява с фактора.
4. $\sqrt{R^2} = |r|$.
5. Знакът пред r се определя от знака на b.

Задължителен изпитен въпрос

Какъв извод може да се направи въз основа на регресионния модел, представен на графиката? Как се интерпретира коефициентът на детерминация в този случай? Каква е стойността на коефициента на корелация?
Очакван отговор за примера от тази глава:

- Всяка 1 седмица увеличение на гестационната възраст се асоциира с 0,24 кг увеличение на очакваното тегло при раждане.
- 76,4 % от вариабилността на теглото при раждане може да се обясни с гестационната възраст.

- $r = \sqrt{0{,}764} = +0{,}87$; знакът е положителен, защото наклонът на правата е положителен.

7.6 Още за регресията

Множествена регресия

Представеният модел съдържа само една обясняваща променлива. В действителност за обяснението на един изход почти винаги се използват няколко предиктора — това е **множествената регресия**:

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k$$

Всеки коефициент b_j се интерпретира „при равни други условия", тоест при непроменени стойности на останалите предиктори. Именно това позволява на множествената регресия да отчита замъгляващи фактори — задача, която в Секция 1 решавахме чрез стандартизация.
При тези модели се среща явлението **колинеарност**: два или повече предиктора са в силна корелация помежду си, при което оценките на техните коефициенти стават нестабилни и трудни за тълкуване.

Когато зависимата променлива е качествена

Ако изходът е дихотомен (оздравял / неоздравял, жив / починал), правата линия не е подходяща: прогнозите ѝ могат да излязат извън интервала от 0 до 1. В такива случаи се използва **логистична регресия**, чиито коефициенти се представят като отношения на шансовете. Тя не е част от настоящия курс, но е най-често срещаният регресионен модел в медицинската литература.

7.7 Чести грешки

Таблица 7.4: Чести грешки при регресионния анализ

#	Грешка	Как да я избегнете
1	Разменени зависима и независима променлива	Y е това, което предсказваме
2	$\sum X^2$ вместо $(\sum X)^2$	Едното е сбор от квадрати, другото — квадрат на сбор
3	Закръгляне на средните или на b преди изчисленията	Води до различни a и прогнози
4	Клинично тълкуване на свободния член a	a е математическа отправна точка
5	Прогноза извън обхвата на данните	Екстраполацията дава числа без смисъл

#	Грешка	Как да я избегнете
6	Съобщаване на r без знак	Знакът идва от наклона b
7	Тълкуване на $\hat{Y}$ като точна индивидуална стойност	$\hat{Y}$ е средно очакваната стойност
8	Извод за причинност от висок R^2	R^2 измерва съответствие, не причина

7.8 Задачи за самостоятелна работа и домашна работа

Задача 7.1

Да се състави модел на зависимостта между възрастта при поставяне на диагноза и преживяемостта при пациенти с белодробен карцином (данните от задача 6.1).

Пациент №	1	2	3	4	5	6	7	8
Възраст X, години	71	71	73	74	75	76	76	78
Преживяемост Y, месеци	135	96	66	44	16	22	11	5

Каква преживяемост се очаква при пациент, диагностициран на 72 години?

Задача 7.2 (домашна работа)

Да се състави модел на зависимостта между изминатото разстояние при 6-минутен тест с ходене и нивото на NT-proBNP (данните от задача 6.2).

Пациент №	1	2	3	4	5	6	7	8
Разстояние X, km	0,7	0,7	0,6	0,5	0,4	0,4	0,3	0,3
NT-proBNP Y, pg/µl	110	120	140	160	180	200	240	300

Интерпретирайте регресионния коефициент с точната му мерна единица.

Задача 7.3 (домашна работа)

Проучване измерва степента на тревожност сред 10 студенти по медицина чрез въпросник със скала от 0 до 100 и записва времето в минути, прекарано в четене по статистика предходния ден.

Време за четене X, мин	45	90	20	120	100	62	70	69	85	15
Тревожност Y, точки	21	84	14	95	84	32	59	43	71	4

а) Постройте регресионния модел. б) Каква тревожност се очаква при студент, който чете 30 минути на ден? в) При $R^2 = 0{,}93$ колко е коефициентът на корелация и какъв е знакът му? г) Може ли да се твърди, че четенето по статистика причинява тревожност?

Задача 7.4

Регресионен модел за връзката между дневния прием на сол (g) и систолното налягане (mmHg) дава $b = 1{,}8$ и $R^2 = 0{,}25$.
а) Как се интерпретира b? б) Колко е коефициентът на корелация и какъв знак има? в) Подходящ ли е този модел за прогнозиране на индивидуално ниво? Обяснете.

Задача 7.5

Два модела предсказват една и съща зависима променлива. Първият има $R^2 = 0{,}81$, вторият — $R^2 = 0{,}36$.
а) Кой модел предсказва по-добре? б) Колко са корелационните коефициенти в двата случая? в) Какво още е необходимо, за да се определи знакът им?

Задача 7.6

Регресионен модел за връзката между възрастта и максималната сърдечна честота дава $\hat{Y} = 208 - 0{,}7 \cdot X$.
а) Интерпретирайте наклона. б) Каква максимална сърдечна честота се очаква при пациент на 60 години? в) Има ли смисъл свободният член в този случай?

Задача 7.7 (за разсъждение)

Изследовател построява регресионен модел върху данни за пациенти на възраст 40–60 години и го използва, за да прогнозира стойност при 15-годишно дете. Каква грешка допуска и как се нарича тя?

Задача 7.8 (за разсъждение)

Модел с $R^2 = 0{,}98$ е построен върху 5 наблюдения. Друг модел със същата зависима променлива има $R^2 = 0{,}45$, но е построен върху 5 000 наблюдения. Кой от двата модела бихте предпочели и защо?

7.9 Кратък тест

За всеки въпрос изберете **един** верен отговор. Ключът е в Секция Г.7.

1. Правата на най-добро съответствие минимизира:

   а) сумата на остатъците;
   б) сумата на квадратите на остатъците;
   в) най-големия остатък;
   г) броя на остатъците.

2. Регресионният коефициент b показва:

   а) стойността на Y при $X = 0$;
   б) дела на обясненото разсейване;
   в) средната промяна в Y при промяна на X с една единица;
   г) силата на връзката в интервала от -1 до $+1$.

3. При $R^2 = 0{,}49$ и отрицателен наклон коефициентът на корелация е:

   а) $+0{,}70$;
   б) $-0{,}70$;
   в) $-0{,}49$;
   г) $-0{,}24$.

4. Свободният член a:

   а) винаги има клинична интерпретация;
   б) е равен на средната стойност на Y;
   в) показва стойността на $\hat{Y}$ при $X = 0$;
   г) е винаги положителен.

5. Използването на модела извън обхвата на наблюдаваните стойности се нарича:

   а) интерполация;
   б) екстраполация;
   в) стандартизация;
   г) детерминация.

6. Кое твърдение за R^2 е вярно?

   а) R^2 може да бъде отрицателен;
   б) R^2 доказва причинна връзка;
   в) R^2 показва дела на обясненото разсейване на зависимата променлива;
   г) R^2 е равен на регресионния коефициент.

8 Колоквиум

Цел на тази глава

Тази глава не въвежда нов материал. Тя обобщава седемте упражнения в един работен алгоритъм и предлага четири примерни варианта с пълни решения, тест за подготовка от 45 въпроса с избираем отговор и 9 отворени въпроса. Част от въпросите изискват понятията от Секция A: видове графики, модели за подбор на извадка, робастност, биномно разпределение и екстраполация.

8.1 Формат и оценяване

Колоквиумът съдържа **две изчислителни задачи**, всяка от които обхваща един от разгледаните методи. Всяка задача се решава изцяло: от изходните данни до словесния извод.

Какво се оценява

1. **Разпознаване на метода** — правилно определен вид на задачата според вида на променливите и въпроса в условието.
2. **Формулиране на хипотезите**, когато задачата е тест на хипотеза.
3. **Записана формула** преди заместването в нея.
4. **Междинни стъпки** — таблиците и изчисленията, а не само крайното число.
5. **Мерна единица** на всяка получена величина.
6. **Словесен извод**, който отговаря на въпроса от условието с имената на изследваните променливи.

Отговор, състоящ се само от число, не носи пълен брой точки дори когато числото е вярно.

Практически съвети

- Носете калкулатор. Формулите се предоставят, но аритметиката е ваша.
- Пишете стъпките така, както са показани в решените примери в наръчника.
- Ако сгрешите в междинна стъпка, но продължите правилно, следващите стъпки се оценяват.
- Оставете си време за последното изречение — изводът носи точки.

8.2 Как се разпознава типът задача

Първата и най-важна стъпка е да се определи какъв е въпросът и какви са променливите. Таблица 8.1 обобщава признаците, по които се разпознава всеки от изучаваните методи.

Таблица 8.1: Разпознаване на типа задача по условието

Ключови думи в условието	Вид на данните	Метод	Глава
„стандартизирани показатели“, „за стандарт да се приеме“	две групи с подгрупова структура	пряка стандартизация	Секция 1
„структура“, „какъв процент от“	едно цяло по категории	екстензивни показатели	Секция 1
„доверителен интервал“, „популационна средна“	една извадка, количествен признак	$\bar{X} \pm t \cdot SE$	Секция 3
„доверителен интервал за относителния дял“	една извадка, качествен признак	$\hat{p} \pm Z \cdot SE_{\hat{p}}$	Секция 3
„има ли разлика“, две групи, средни стойности	количествен признак, две независими групи	t-тест	Секция 4
„има ли разлика“, две групи, проценти или брой от брой	качествен признак, две независими групи	Z-тест за два дяла	Секция 4
„има ли зависимост / връзка“, таблица с честоти	два качествени признака	χ^2-тест, φ или V	Секция 5
„силата и посоката на зависимостта“	два количествени признака	коефициент на Pearson	Секция 6
„да се състави модел“, „да се моделира“, „да се прогнозира“	два количествени признака	линейна регресия	Секция 7

⚠ Двете най-чести грешки при разпознаването

1. **„Има ли разлика“ срещу „има ли зависимост“.** Първото сравнява две групи по един признак (t- или Z-тест). Второто проверява връзка между два признака (χ^2-тест или корелация).
2. **Корелация срещу регресия.** „Да се измери силата и посоката“ означава корелация. „Да се състави модел“ или „да се прогнозира“ означава регресия. Двете задачи могат да са построени върху едни и същи данни.

8.3 Общ ход на решението

Независимо от метода, решението следва един и същ скелет:

1. **Определяне на типа задача** и на вида на променливите.
2. **Хипотези** (H_0 и H_1) и **ниво на значимост** α — при всички тестове на хипотеза.
3. **Описателни изчисления** — средни, дялове, суми, помощни таблици.
4. **Формула** — записана, преди да се замести в нея.
5. **Тестова статистика** или **интервал**.
6. **Степени на свобода** и **p-стойност** по таблица.
7. **Извод с думи**, отговарящ на въпроса от условието.
8. **Сила на връзката**, когато методът я изисква (V, φ, r, R^2).

8.4 Четири примерни варианта

8.4.1 Вариант 1

Задача 1

Да се изчислят стандартизираните показатели за леталитет в две областни болници. За стандарт да се приеме структурата по отделения в **болница Б**.

Отделение	Преминали (А)	Починали (А)	Преминали (Б)	Починали (Б)
Инфекциозни болести	500	10	1 500	35
Хирургия	1 500	50	2 000	70
Онкология	1 000	85	1 500	120
Общо	**3 000**	**145**	**5 000**	**225**

Задача 2

Ретроспективно проучване анализира наблюдаваните усложнения след аборт, извършен по два различни метода. Има ли зависимост между метода за аборт и вида на усложнението?

Метод за аборт	Кръвотечение	Инфекция	Инфертилитет	Общо
Класически	10	26	24	60
Аспирационен	12	13	15	40
Общо	**22**	**39**	**39**	**100**

8.4.2 Вариант 2

Задача 1

Да се изчисли 95 % доверителен интервал за средния популационен инкубационен период при пациенти с хепатит А.

Инкубационен период, дни	Честота f, брой пациенти
1 – 15	40
16 – 30	120
31 – 45	75
46 – 60	14
61 – 75	1
Общо	**250**

Задача 2

Използвайки данните от таблицата, да се измери силата и посоката на зависимостта между възрастта при поставяне на диагноза и преживяемостта при пациенти с белодробен карцином.

Пациент №	1	2	3	4	5	6	7	8
Възраст X, години	71	71	73	74	75	76	76	78
Преживяемост Y, месеци	135	96	66	44	16	22	11	5

8.4.3 Вариант 3

Задача 1

400 пациенти с множествена склероза са включени в клинично проучване. 25 от 200 лекувани с експериментална терапия рецидивират в рамките на 6 месеца след края на терапевтичния курс, докато при конвенционалната терапия този брой е 45 от 200. Има ли статистически значима разлика между двата вида терапия по отношение на относителния дял на рецидивиращите пациенти?

Задача 2

В таблицата са отчетени изминатото разстояние при 6-минутен тест с ходене (6MWT) и нивото на NT-proBNP при 8 пациенти с начална сърдечна недостатъчност. Да се състави модел на зависимостта между изминатото разстояние и нивото на NT-proBNP.

Пациент №	1	2	3	4	5	6	7	8
Разстояние X, km	0,7	0,7	0,6	0,5	0,4	0,4	0,3	0,3
NT-proBNP Y, pg/µl	110	120	140	160	180	200	240	300

8.4.4 Вариант 4

Задача 1

Пациенти с недребноклетъчен белодробен карцином са разпределени в две независими групи според приложената терапия. Има ли разлика между двете групи по отношение на средната преживяемост?

Вид лечение	Обем на извадката	Средна преживяемост	Стандартно отклонение
Експериментално	$n_1 = 40$	$\bar{X}_1 = 18{,}5$ месеца	$S_1 = 5{,}2$ месеца

Конвенционално	$n_2 = 40$	$\bar{X}_2 = 15{,}0$ месеца	$S_2 = 4{,}0$ месеца

Задача 2

Проучване при 300 пациенти с диабет тип 2 регистрира придържането към терапията и постигането на прицелна стойност на HbA1c. Има ли зависимост между придържането към терапията и постигането на прицелната стойност?

Придържане към терапията	Прицелен HbA1c	Без прицелен HbA1c	Общо
Добро	96	84	180
Лошо	36	84	120
Общо	**132**	**168**	**300**

8.5 Решения на вариантите

8.5.1 Решение на вариант 1, задача 1 (стандартизация)

Стъпка 1. Нестандартизирани показатели за леталитет.

$$IR = \frac{\text{починали}}{\text{преминали}} \times 100$$

Таблица 8.2: Нестандартизирани показатели

Отделение	IR_A, %	IR_B, %
Инфекциозни болести	$10/500 \times 100 = 2{,}00$	$35/1\,500 \times 100 = 2{,}33$
Хирургия	$50/1\,500 \times 100 = 3{,}33$	$70/2\,000 \times 100 = 3{,}50$
Онкология	$85/1\,000 \times 100 = 8{,}50$	$120/1\,500 \times 100 = 8{,}00$
Общо	$145/3\,000 \times 100 = \mathbf{4{,}83}$	$225/5\,000 \times 100 = \mathbf{4{,}50}$

Погледнати общо, болница А има **по-висок** леталитет (4,83 % срещу 4,50 %), въпреки че по отделения разликите са малки и разнопосочни.

Стъпка 2. Стандарт — структурата по отделения в болница Б.

$$S = \frac{\text{преминали в отделението (Б)}}{\text{всички преминали (Б)}} \times 100$$

Таблица 8.3: Изчисляване на стандарта

Отделение	Преминали (Б)	S, %
Инфекциозни болести	1 500	30,0
Хирургия	2 000	40,0
Онкология	1 500	30,0
Общо	**5 000**	**100,0**

Стъпка 3. Стандартизирани показатели.

$$AIR = \frac{IR \times S}{100}$$

Таблица 8.4: Стандартизирани показатели

Отделение	S, %	AIR_A, %	AIR_B, %
Инфекциозни болести	30,0	$2{,}00 \cdot 30/100 = 0{,}60$	$2{,}33 \cdot 30/100 = 0{,}70$
Хирургия	40,0	$3{,}33 \cdot 40/100 = 1{,}33$	$3{,}50 \cdot 40/100 = 1{,}40$
Онкология	30,0	$8{,}50 \cdot 30/100 = 2{,}55$	$8{,}00 \cdot 30/100 = 2{,}40$

Отделение	S, %	AIR_A, %	AIR_B, %
Общо	**100,0**	**4,48**	**4,50**

Стъпка 4. Заключение. Стандартизираните показатели за леталитет са **4,48 %** за болница А и **4,50 %** за болница Б. След уеднаквяване на структурата по отделения разликата между двете болници практически изчезва. По-високият нестандартизиран показател в болница А се дължи на различния състав на преминалите пациенти. В болница А делът на онкологичните пациенти, при които леталитетът е най-висок, е 33,3 % срещу 30,0 % в болница Б, а делът на инфекциозно болните, при които леталитетът е най-нисък, е само 16,7 % срещу 30,0 % в болница Б.

Проверка: болница Б, чиято структура е приета за стандарт, запазва общия си показател от 4,50 % преди и след стандартизацията.

8.5.2 Решение на вариант 1, задача 2 (χ^2-тест)

Стъпка 1. Хипотези. H_0: няма зависимост между метода за аборт и вида на усложнението. H_1: има зависимост. Ниво на значимост $\alpha = 0{,}05$.

Стъпка 2. Очаквани честоти.

$$E_{ij} = \frac{R_i \times C_j}{N}$$

$$E_{11} = \frac{60 \times 22}{100} = 13{,}2 \qquad E_{12} = \frac{60 \times 39}{100} = 23{,}4 \qquad E_{13} = \frac{60 \times 39}{100} = 23{,}4$$

$$E_{21} = \frac{40 \times 22}{100} = 8{,}8 \qquad E_{22} = \frac{40 \times 39}{100} = 15{,}6 \qquad E_{23} = \frac{40 \times 39}{100} = 15{,}6$$

Всички очаквани честоти са ≥ 5 — условията за прилагане са изпълнени.

Стъпка 3. Тестова статистика.

Таблица 8.5: Изчисляване на χ^2

Клетка	O	E	$(O-E)^2/E$
Класически × кръвотечение	10	13,2	0,776
Класически × инфекция	26	23,4	0,289
Класически × инфертилитет	24	23,4	0,015
Аспирационен × кръвотечение	12	8,8	1,164
Аспирационен × инфекция	13	15,6	0,433
Аспирационен × инфертилитет	15	15,6	0,023
Общо	**100**	**100**	**2,70**

$$\chi^2 = 0{,}776 + 0{,}289 + 0{,}015 + 1{,}164 + 0{,}433 + 0{,}023 = 2{,}70$$

Стъпка 4. Степени на свобода и p-стойност.

$$df = (R-1)(C-1) = (2-1)(3-1) = 2$$

Критичната стойност при $df = 2$ и $p = 0{,}10$ е 4,605. Понеже $2{,}70 < 4{,}605$, следва $p > 0{,}10$.

Стъпка 5. Сила на връзката. $k = \min(2,\ 3) = 2$, тоест $k - 1 = 1$:

$$V = \sqrt{\frac{2{,}70}{100 \cdot 1}} = \sqrt{0{,}027} = 0{,}16$$

Стъпка 6. Заключение. $p > 0{,}10 > \alpha$. Не отхвърляме H_0. Данните не дават достатъчно доказателства за зависимост между метода за аборт и вида на усложнението ($\chi^2 = 2{,}70$; $df = 2$; $p > 0{,}10$; $V = 0{,}16$). Липсата на статистическа значимост при $N = 100$ не доказва, че зависимост няма — извадката е малка и мощността е ниска.

8.5.3 Решение на вариант 2, задача 1 (доверителен интервал)

Стъпка 1. Среди на интервалите и средна аритметична.

Таблица 8.6: Работна таблица

Период, дни	f	X_u	$X_u \cdot f$	$X_u - \bar{X}$	$(X_u - \bar{X})^2 \cdot f$
1 – 15	40	8	320	−18,96	14 379,26
16 – 30	120	23	2 760	−3,96	1 881,79
31 – 45	75	38	2 850	11,04	9 141,12
46 – 60	14	53	742	26,04	9 493,14
61 – 75	1	68	68	41,04	1 684,28
Общо	**250**		**6 740**		**36 579,60**

$$\bar{X} = \frac{\sum(X_u \cdot f)}{\sum f} = \frac{6\,740}{250} = 26{,}96 \text{ дни}$$

Стъпка 2. Стандартно отклонение.

$$S = \sqrt{\frac{\sum(X_u - \bar{X})^2 \cdot f}{n-1}} = \sqrt{\frac{36\,579{,}60}{249}} = \sqrt{146{,}91} = 12{,}12 \text{ дни}$$

Стъпка 3. Стандартна грешка на средната.

$$SE_{\bar{X}} = \frac{S}{\sqrt{n}} = \frac{12{,}12}{\sqrt{250}} = \frac{12{,}12}{15{,}81} = 0{,}77 \text{ дни}$$

Стъпка 4. Доверителен интервал. При $n = 250$ критичната стойност на t-разпределението практически съвпада с $Z = 1{,}96$:

$$95\ \%\ CI = 26{,}96 \pm 1{,}96 \cdot 0{,}77 = 26{,}96 \pm 1{,}51$$

- долна граница: 25,45 дни;
- горна граница: 28,47 дни.

Стъпка 5. Заключение. С 95 % гаранционна вероятност средният инкубационен период при пациенти с хепатит А в популацията е между **25,45 и 28,47 дни**.

8.5.4 Решение на вариант 2, задача 2 (корелация)

Стъпка 1. Средни аритметични.

$$\bar{X} = \frac{594}{8} = 74{,}25 \text{ години}, \qquad \bar{Y} = \frac{395}{8} = 49{,}375 \text{ месеца}$$

Стъпка 2. Отклонения, кръстосани продукти и квадрати.

Таблица 8.7: Работна таблица за корелацията

№	X	Y	d_x	d_y	$d_x \cdot d_y$	d_x^2	d_y^2
1	71	135	−3,25	85,625	−278,28	10,5625	7 331,64
2	71	96	−3,25	46,625	−151,53	10,5625	2 173,89
3	73	66	−1,25	16,625	−20,78	1,5625	276,39
4	74	44	−0,25	−5,375	1,34	0,0625	28,89
5	75	16	0,75	−33,375	−25,03	0,5625	1 113,89
6	76	22	1,75	−27,375	−47,91	3,0625	749,39
7	76	11	1,75	−38,375	−67,16	3,0625	1 472,64
8	78	5	3,75	−44,375	−166,41	14,0625	1 969,14
Общо	**594**	**395**	**0,00**	**0,00**	**−755,75**	**43,50**	**15 115,88**

Стъпка 3. Коефициент на Pearson.

$$r = \frac{\sum (d_x \cdot d_y)}{\sqrt{\sum d_x^2 \cdot \sum d_y^2}} = \frac{-755{,}75}{\sqrt{43{,}50 \cdot 15\,115{,}88}} = \frac{-755{,}75}{\sqrt{657\,540{,}8}} = \frac{-755{,}75}{810{,}89} = -0{,}93$$

Стъпка 4. Статистическа значимост.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0{,}93 \cdot \sqrt{6}}{\sqrt{1-0{,}8649}} = \frac{-2{,}278}{0{,}3676} = -6{,}20$$

$$df = n - 2 = 6$$

Критичната стойност при $df = 6$ и $p = 0{,}01$ е 3,707. Понеже $|-6{,}20| > 3{,}707$, следва $p < 0{,}01$.

Стъпка 5. Заключение. $r = -0{,}93$. Съществува **силна обратна** връзка между възрастта при поставяне на диагноза и преживяемостта при пациенти с белодробен карцином: по-високата възраст при диагностициране се свързва с по-кратка преживяемост. Връзката е статистически значима ($t = -6{,}20$; $df = 6$; $p < 0{,}01$). Изводът важи за наблюдавания възрастов диапазон 71–78 години и не доказва причинна зависимост.

8.5.5 Решение на вариант 3, задача 1 (Z-тест за два дяла)

Стъпка 1. Хипотези. H_0: няма разлика между популационните дялове рецидивиращи пациенти при двете терапии. H_1: има разлика. $\alpha = 0{,}05$.

Стъпка 2. Извадкови дялове.

$$\hat{p}_1 = \frac{25}{200} \times 100 = 12{,}5\,\% \qquad \hat{p}_2 = \frac{45}{200} \times 100 = 22{,}5\,\%$$

Клиничният ефект е $\Delta = 22{,}5 - 12{,}5 = 10$ процентни пункта.

Стъпка 3. Обединен дял.

$$\hat{p} = \frac{25 + 45}{200 + 200} \times 100 = \frac{70}{400} \times 100 = 17{,}5\,\%$$

Стъпка 4. Тестова статистика.

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(100 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{12{,}5 - 22{,}5}{\sqrt{17{,}5 \cdot 82{,}5 \cdot \left(\frac{1}{200} + \frac{1}{200}\right)}}$$

$$Z = \frac{-10}{\sqrt{1\,443{,}75 \cdot 0{,}01}} = \frac{-10}{\sqrt{14{,}44}} = \frac{-10}{3{,}80} = -2{,}63$$

Стъпка 5. p-стойност. $|Z| = 2{,}63 > 2{,}576$, следователно $p < 0{,}01$ (софтуерът дава $p = 0{,}008$).

Стъпка 6. Заключение. $p < 0{,}01 < \alpha$. Отхвърляме H_0 и приемаме H_1. Има статистически значима разлика между двата вида терапия по отношение на относителния дял рецидивиращи пациенти. При експерименталната терапия делът на рецидивиралите е 12,5 % срещу 22,5 % при конвенционалната, тоест с 10 процентни пункта по-нисък.

8.5.6 Решение на вариант 3, задача 2 (регресия)

Стъпка 1. Определяне на променливите. Изминатото разстояние X е независимата (факторна) променлива, а нивото на NT-proBNP Y — зависимата (резултативна).

Стъпка 2. Помощна таблица.

Таблица 8.8: Помощна таблица

№	X	Y	$X \cdot Y$	X^2
1	0,7	110	77,0	0,49
2	0,7	120	84,0	0,49
3	0,6	140	84,0	0,36
4	0,5	160	80,0	0,25
5	0,4	180	72,0	0,16
6	0,4	200	80,0	0,16
7	0,3	240	72,0	0,09
8	0,3	300	90,0	0,09
$n = 8$	**3,9**	**1 450**	**639,0**	**2,09**

Стъпка 3. Регресионен коефициент.

$$b = \frac{n\sum(XY) - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2} = \frac{8 \cdot 639 - 3{,}9 \cdot 1\,450}{8 \cdot 2{,}09 - 3{,}9^2}$$

$$b = \frac{5\,112 - 5\,655}{16{,}72 - 15{,}21} = \frac{-543}{1{,}51} = -359{,}60$$

Стъпка 4. Свободен член.

$$\bar{X} = \frac{3{,}9}{8} = 0{,}4875 \text{ км}, \qquad \bar{Y} = \frac{1\,450}{8} = 181{,}25 \text{ pg/µl}$$

$$a = \bar{Y} - b \cdot \bar{X} = 181{,}25 - (-359{,}60 \cdot 0{,}4875) = 181{,}25 + 175{,}31 = 356{,}56$$

Стъпка 5. Регресионно уравнение.

$$\hat{Y} = -359{,}60 \cdot X + 356{,}56$$

Стъпка 6. Прогноза. За пациент, изминал 0,55 км:

$$\hat{Y} = -359{,}60 \cdot 0{,}55 + 356{,}56 = -197{,}78 + 356{,}56 = 158{,}78 \text{ pg/µl}$$

Стъпка 7. Заключение. Всеки допълнителен изминат километър при 6-минутния тест с ходене се асоциира със средно **360 pg/µl по-ниско** ниво на NT-proBNP. Изразено в по-практична мерна единица: всеки допълнителни 100 метра съответстват на около 36 pg/µl по-ниска стойност. При $R^2 = 0{,}845$ 84,5 % от вариабилността на NT-proBNP се обяснява с изминатото разстояние, а коефициентът на корелация е $r = -\sqrt{0{,}845} = -0{,}92$; знакът е отрицателен, защото наклонът на правата е отрицателен.

8.5.7 Решение на вариант 4, задача 1 (t-тест)

Стъпка 1. Хипотези. H_0: няма разлика между двете популационни средни преживяемости. H_1: има разлика. $\alpha = 0{,}05$.

Стъпка 2. Клиничен ефект.

$$\Delta = \bar{X}_1 - \bar{X}_2 = 18{,}5 - 15{,}0 = 3{,}5 \text{ месеца}$$

Стъпка 3. Тестова статистика.

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{|18{,}5 - 15{,}0|}{\sqrt{\frac{5{,}2^2}{40} + \frac{4{,}0^2}{40}}}$$

$$t = \frac{3{,}5}{\sqrt{\frac{27{,}04}{40} + \frac{16{,}00}{40}}} = \frac{3{,}5}{\sqrt{0{,}676 + 0{,}400}} = \frac{3{,}5}{\sqrt{1{,}076}} = \frac{3{,}5}{1{,}04} = 3{,}37$$

Стъпка 4. Степени на свобода и p-стойност.

$$df \approx \min(n_1 - 1,\ n_2 - 1) = \min(39,\ 39) = 39$$

Критичната стойност при $df = 39$ и $p = 0{,}01$ е 2,708. Понеже $3{,}37 > 2{,}708$, следва $p < 0{,}01$ (софтуерът дава $p = 0{,}001$).

Стъпка 5. Заключение. $p < 0{,}01 < \alpha$. Отхвърляме H_0 и приемаме H_1. Има статистически значима разлика между двете групи по отношение на средната преживяемост: при експерименталното лечение тя е с 3,5 месеца по-висока (18,5 срещу 15,0 месеца). Дали разликата е клинично значима, се преценява отделно, спрямо предварително зададена клинично важна разлика и спрямо профила на нежеланите реакции.

8.5.8 Решение на вариант 4, задача 2 (χ^2-тест, таблица 2 × 2)

Стъпка 1. Хипотези. H_0: няма зависимост между придържането към терапията и постигането на прицелен HbA1c. H_1: има зависимост. $\alpha = 0{,}05$.

Стъпка 2. Описание. Прицелна стойност постигат 96 от 180 пациенти с добро придържане (53,3 %) и 36 от 120 с лошо придържане (30,0 %).

Стъпка 3. Очаквани честоти.

$$E_{11} = \frac{180 \times 132}{300} = 79{,}2 \qquad E_{12} = \frac{180 \times 168}{300} = 100{,}8$$

$$E_{21} = \frac{120 \times 132}{300} = 52{,}8 \qquad E_{22} = \frac{120 \times 168}{300} = 67{,}2$$

Всички очаквани честоти са ≥ 5.

Стъпка 4. Тестова статистика.

$$\chi^2 = \frac{(96 - 79{,}2)^2}{79{,}2} + \frac{(84 - 100{,}8)^2}{100{,}8} + \frac{(36 - 52{,}8)^2}{52{,}8} + \frac{(84 - 67{,}2)^2}{67{,}2}$$

$$\chi^2 = 3{,}56 + 2{,}80 + 5{,}35 + 4{,}20 = 15{,}91$$

Стъпка 5. Степени на свобода и p-стойност.

$$df = (2 - 1)(2 - 1) = 1$$

Критичната стойност при $df = 1$ и $p = 0{,}01$ е 6,635. Понеже $15{,}91 > 6{,}635$, следва $p < 0{,}01$ (софтуерът дава $p < 0{,}001$).

Стъпка 6. Сила и посока на връзката.

$$\varphi = \frac{(96 \cdot 84) - (36 \cdot 84)}{\sqrt{180 \cdot 120 \cdot 132 \cdot 168}} = \frac{8\,064 - 3\,024}{\sqrt{479\,001\,600}} = \frac{5\,040}{21\,886{,}1} = 0{,}23$$

Стъпка 7. Заключение. Има статистически значима, но слаба до умерена по сила зависимост между придържането към терапията и постигането на прицелен HbA1c ($\chi^2 = 15{,}91$; $df = 1$; $p < 0{,}001$; $\varphi = 0{,}23$). Положителният знак показва, че доброто придържане се свързва с по-често постигане на прицелната стойност.

8.6 Тест за подготовка, част 1

Първата част обхваща материала от седемте упражнения. За всеки въпрос изберете **един** верен отговор. Ключът и кратките обяснения са в Секция Г.8.

1. Показателят „35 % от преминалите през отделението пациенти са приети по спешност“ е:

   а) абсолютна величина;
   б) екстензивен показател;
   в) интензивен показател;
   г) стандартизиран показател.

2. След стандартизация групата, чиято структура е приета за стандарт:

   а) винаги получава по-нисък показател;
   б) винаги получава по-висок показател;
   в) запазва общия си показател непроменен;
   г) получава показател, равен на 100 %.

3. При силно дясно изтеглено разпределение на болничния престой най-подходящо е да се съобщят:

   а) средна аритметична и стандартно отклонение;
   б) медиана и интерквартилен размах;
   в) мода и размах;
   г) средна аритметична и размах.

4. Извадка от 400 новородени има средно тегло 3 400 г и стандартно отклонение 500 г. Приблизително колко новородени тежат между 2 400 г и 4 400 г?

   а) 272;
   б) 320;
   в) 380;
   г) 399.

5. Стандартната грешка на средната аритметична:

   а) описва разсейването на отделните наблюдения;
   б) нараства с увеличаване на обема на извадката;
   в) намалява пропорционално на $\sqrt{n}$;
   г) е винаги по-голяма от стандартното отклонение.

6. 95 % доверителен интервал за средна преживяемост е [6,6; 9,4] месеца. Коя интерпретация е правилна?

   а) 95 % от пациентите преживяват между 6,6 и 9,4 месеца;
   б) Има точно 95 % вероятност истинската средна да е между тези граници;
   в) При многократно повторение на процедурата около 95 % от построените интервали ще съдържат истинската средна;
   г) Средната в извадката е между 6,6 и 9,4 месеца с вероятност 95 %.

7. Изследовател намалява нивото на значимост от $\alpha = 0{,}05$ на $\alpha = 0{,}01$ при непроменени обем на извадката и размер на ефекта. Мощността на теста:

   а) нараства;
   б) намалява;
   в) остава непроменена;
   г) става равна на β.

8. Пациенти са изследвани преди и след 12-седмична рехабилитация по отношение на нормално разпределен количествен показател. Подходящият тест е:

   а) t-тест за две независими извадки;
   б) t-тест за свързани извадки;
   в) Z-тест за два относителни дяла;
   г) χ^2-тест.

9. Резултатът от проучване е $t = 1{,}42$ при $df = 60$. Следва, че:

   а) $p < 0{,}01$ и отхвърляме H_0;
   б) $p < 0{,}05$ и отхвърляме H_0;
   в) $p > 0{,}10$ и не отхвърляме H_0;
   г) p-стойността не може да се определи без α.

10. За таблица на съчетаемост 3 × 4 степените на свобода са:

    а) 12;
    б) 7;
    в) 6;
    г) 5.

11. Кое от следните **не** е условие за прилагане на χ^2-теста?

    а) Наблюденията са независими;
    б) В клетките стоят броеве, а не проценти;
    в) Изследваната променлива е нормално разпределена;
    г) Поне 80 % от очакваните честоти са ≥ 5.

12. Коефициентът V на Cramér се съобщава заедно с χ^2, защото:

    а) заменя p-стойността;
    б) измерва силата на връзката независимо от обема на извадката;
    в) доказва причинна зависимост;
    г) показва посоката на връзката при таблица 3 × 3.

13. Диаграмата на разсейване показва ясна U-образна зависимост между два количествени признака, а изчисленият коефициент на Pearson е $r = 0{,}03$. Правилният извод е:

    а) между двете променливи няма никаква връзка;
    б) между двете променливи няма линейна връзка, но има нелинейна;
    в) данните съдържат грешка при въвеждането;
    г) необходима е по-голяма извадка.

14. Регресионен модел дава $\hat{Y} = 208 - 0{,}7 \cdot X$, където X е възрастта в години, а Y – максималната сърдечна честота. Коефициентът $-0{,}7$ означава, че:

    а) 70 % от вариабилността се обяснява с възрастта;
    б) корелационният коефициент е $-0{,}7$;
    в) при увеличаване на възрастта с 1 година очакваната максимална сърдечна честота намалява средно със 0,7 удара в минута;
    г) максималната сърдечна честота при раждане е 0,7 удара в минута.

15. Регресионен модел има $R^2 = 0{,}64$, а наклонът на правата е отрицателен. Коефициентът на корелация е:

    а) $+0{,}80$;
    б) $-0{,}80$;
    в) $-0{,}64$;
    г) $-0{,}41$.

8.7 Тест за подготовка, част 2

Втората част включва и понятия от Секция А: видове графики, модели за подбор на извадка, робастност и екстраполация. За всеки въпрос изберете **един** верен отговор.

16. Какъв вид графика се препоръчва за визуализиране на разпределението на една качествена променлива, измерена на **номинална** скала?

    а) кръгова диаграма;
    б) стълбовидна диаграма;
    в) хистограма;
    г) диаграма кутия (боксплот).

17. Общият **нестандартизиран** показател е:

    а) сума от нестандартизираните резултати по подгрупи;
    б) сума от стандартизираните резултати по подгрупи;
    в) произведение на нестандартизираните резултати по подгрупи;
    г) всички отговори са грешни.

18. Променливата „диагноза“, дефинирана като „рак на белия дроб, рак на гърдата, рак на дебелото черво или друго“, се измерва на скала:

    а) номинална;
    б) ординална;
    в) интервална;
    г) пропорционална.

19. Свободен член $a = 1{,}85$ в регресионния модел означава, че:

    а) наклонът на правата на най-добро съответствие може да бъде както положителен, така и отрицателен;
    б) наклонът на правата е положителен;
    в) наклонът на правата е отрицателен;
    г) всички отговори са грешни.

20. Пациенти с множествена склероза са разделени на 2 групи според терапията. 95 % CI за честотата на рецидиви при терапия А е между 21 % и 28 %, а при терапия Б — между 32 % и 42 %. Какъв извод може да се направи?

    а) най-вероятно няма разлика в ефективността на двете терапии;
    б) най-вероятно терапия А е статистически значимо по-ефективна от терапия Б;
    в) най-вероятно терапия Б е статистически значимо по-ефективна от терапия А;
    г) всички изводи са грешни.

21. Кой отговор подрежда правилно моделите за подбор на извадка според очаквания размер на **случайната** грешка — от най-малка към най-голяма?

    а) типологична извадка, квотна извадка, клъстерна извадка;
    б) клъстерна извадка, квотна извадка, типологична извадка;
    в) извадка по удобство, стратифицирана извадка, квотна извадка;
    г) стратифицирана извадка, извадка по удобство, типологична извадка.

22. Пациенти с белодробен карцином са разделени на 2 групи според терапията. Честотата на рецидиви при терапия А е 18 %, а при терапия Б — 24 % ($p = 0{,}050$). Какъв извод може да се направи?

    а) няма основание да се твърди, че двете терапии се различават;
    б) терапия А е статистически значимо по-ефективна от терапия Б;
    в) терапия Б е статистически значимо по-ефективна от терапия А;
    г) всички изводи са грешни.

23. Наличието на аутлайъри в данните ще окаже **най-слабо** въздействие върху кой показател?

    а) стандартна грешка на средната аритметична;
    б) средна аритметична;
    в) медиана;
    г) стандартно отклонение.

24. В проучване на родилки с множествена склероза средното тегло на новородените е 2 700 г при майките в ремисия над 1 година и 2 100 г при майките в ремисия под 1 година ($p = 0{,}006$); използван е t-тест за две независими извадки. Кое твърдение е **вярно**?

    а) резултатът е статистически значим при $\alpha = 0{,}05$ и нулевата хипотеза може да бъде отхвърлена;
    б) тестът на McNemar също може да се използва в този случай;
    в) теглото на новородените най-вероятно не е нормално разпределена величина;
    г) алтернативната хипотеза гласи, че средното тегло е еднакво в двете групи.

25. Кое от следните твърдения е **грешно**?

    а) интерполацията е прогнозиране в рамките на областта от известни стойности, послужили за построяване на модела;
    б) екстраполацията е прогнозиране извън тази област;
    в) чрез екстраполация се постига по-голяма валидност на получената оценка;
    г) екстраполацията поставя условие наблюдаваната зависимост да се запази и извън областта от известните стойности.

26. Параметричните тестове за проверка на хипотеза **губят** робастност при:

    а) малки по обем извадки;
    б) отсъствие на аутлайъри;
    в) приблизително еднакви по обем и разсейване извадки;
    г) всички отговори са верни.

27. При какво условие t-разпределението се доближава по стойности до Z-разпределението?

    а) при нормално разпределени величини;
    б) при големи по обем извадки;
    в) при количествени променливи, измерени на пропорционална скала;
    г) при дизайн със свързани извадки.

28. В проучване постоперативни усложнения са наблюдавани при 23 (10 %) от общо 230 пациенти; 95 % CI за P е между 6 % и 14 %. Кое твърдение е **грешно**?

    а) В популацията честотата на постоперативни усложнения най-вероятно е по-голяма от 6 %;
    б) В изследваната извадка честотата на постоперативните усложнения може да приеме стойност между 6 % и 14 %;
    в) авторите биха получили оценка с по-голямо ниво на доверие, ако бяха изчислили 99 % CI;
    г) авторите биха получили по-прецизна оценка, ако бяха включили повече пациенти.

29. При равни други условия **най-малък** брой единици на наблюдение ще бъде необходим при:

    а) разнородна популация и високо ниво на гаранционна вероятност;
    б) еднородна популация и високо ниво на гаранционна вероятност;
    в) разнородна популация и ниско ниво на гаранционна вероятност;
    г) еднородна популация и ниско ниво на гаранционна вероятност.

30. В проучване със 121 проследявани пациенти средната преживяемост е 44 месеца със стандартно отклонение 22 месеца. На колко е равна стандартната грешка на средната аритметична?

    а) 6 месеца;
    б) 4 месеца;
    в) 2 месеца;
    г) 1 месец.

8.8 Тест за подготовка, част 3

31. В проучване на родилки с множествена склероза средното тегло на новородените е 2 700 г при майките в ремисия над 1 година и 2 100 г при майките в ремисия под 1 година ($p = 0{,}006$); използван е t-тест за две независими извадки. Кое твърдение е **грешно**?

    а) резултатът е статистически значим при $\alpha = 0{,}05$ и нулевата хипотеза може да бъде отхвърлена;
    б) тестът на McNemar също може да се използва в този случай;
    в) теглото на новородените е най-вероятно нормално разпределена величина;
    г) работната хипотеза гласи, че средното тегло е еднакво в двете групи.

32. За използване на критичните стойности от Z-разпределението са необходими достатъчно голяма извадка и надеждна оценка на кой популационен параметър?

    а) средна аритметична стойност;
    б) стандартна грешка;
    в) стандартно отклонение;
    г) ексцес.

33. При равни други условия **най-голям** брой единици на наблюдение ще бъде необходим при:

    а) разнородна популация и високо ниво на гаранционна вероятност;
    б) еднородна популация и високо ниво на гаранционна вероятност;
    в) разнородна популация и ниско ниво на гаранционна вероятност;
    г) еднородна популация и ниско ниво на гаранционна вероятност.

34. Променливата „ръст в см“ се измерва на скала:

    а) номинална;
    б) ординална;
    в) интервална;
    г) пропорционална.

35. Пациенти с множествена склероза са разделени на 2 групи според терапията. 95 % CI за честотата на рецидиви при терапия А е между 21 % и 28 %, а при терапия Б — между 24 % и 38 %. Какъв извод може да се направи?

    а) най-вероятно терапия А е статистически значимо по-ефективна от терапия Б;
    б) най-вероятно терапия Б е статистически значимо по-ефективна от терапия А;
    в) данните не позволяват извод за разлика между двете терапии;
    г) двете терапии са доказано еднакво ефективни.

36. Кой отговор подрежда правилно моделите за подбор на извадка според очаквания размер на **случайната** грешка — от най-голяма към най-малка?

    а) типологична извадка, квотна извадка, клъстерна извадка;
    б) клъстерна извадка, квотна извадка, типологична извадка;
    в) извадка по удобство, стратифицирана извадка, квотна извадка;
    г) стратифицирана извадка, извадка по удобство, типологична извадка.

37. Пациенти с белодробен карцином са разделени на 2 групи според терапията. Честотата на рецидиви при терапия А е 28 %, а при терапия Б — 24 % ($p = 0{,}0051$). Какъв извод може да се направи?

    а) терапия А е статистически значимо по-ефективна от терапия Б;
    б) терапия Б е статистически значимо по-ефективна от терапия А;
    в) няма разлика в ефективността на двете терапии;
    г) всички изводи са грешни.

38. Регресионен коефициент $b = -1{,}85$ означава, че:

    а) наклонът на правата може да бъде както положителен, така и отрицателен;
    б) наклонът на правата е положителен;
    в) наклонът на правата е отрицателен;
    г) всички отговори са грешни.

39. Кое от следните твърдения е **вярно**?

    а) ординалната скала притежава най-висока точност на измерване;
    б) при ординалната скала разстоянието между категориите е еднакво за всяка част от скалата;
    в) при ординалната скала една единица може да принадлежи към повече от една категория;
    г) при ординалната скала различните категории имат логическа подредба.

40. Наличието на аутлайъри в данните ще окаже **най-силно** въздействие върху кой показател?

    а) стандартно отклонение;
    б) интерквартилен размах;
    в) мода;
    г) ниво на значимост.

41. Общият **стандартизиран** показател е:

    а) произведение на стандартизираните резултати по подгрупи;
    б) сума от нестандартизираните резултати по подгрупи;
    в) сума от стандартизираните резултати по подгрупи;
    г) всички отговори са грешни.

42. В проучване с 81 пациенти средната преживяемост е 36 месеца със стандартно отклонение 9 месеца. На колко е равна стандартната грешка на средната аритметична?

    а) 9 месеца;
    б) 4 месеца;
    в) 2 месеца;
    г) 1 месец.

43. Параметричните тестове за проверка на хипотеза **запазват** робастност при:

    а) малки по обем извадки;
    б) наличие на аутлайъри;

в) приблизително еднакви по обем и разсейване извадки;
г) всички отговори са грешни.

44. В проучване преддиабетно състояние е установено при 27 от 91 деца (29,7 %); 95 % CI за P е между 20,3 % и 39,1 %. Кое твърдение е **грешно**?

    а) в популацията от деца честотата на преддиабет най-вероятно е между 20,3 % и 39,1 %;
    б) вероятността произволно дете да бъде в преддиабетно състояние е между 20,3 % и 39,1 %;
    в) в изследваната извадка честотата на преддиабет е 29,7 %;
    г) авторите биха получили оценка с по-голяма гаранционна вероятност, ако бяха изчислили 99 % CI.

45. Какъв вид графика се препоръчва за визуализиране на разпределението на една качествена променлива, измерена на **ординална** скала?

    а) кръгова диаграма;
    б) стълбовидна диаграма;
    в) хистограма;
    г) диаграма кутия (боксплот).

8.9 Отворени въпроси

Отворените въпроси изискват кратък писмен отговор, а не избор между готови твърдения. Очаква се обяснение с термините от курса.

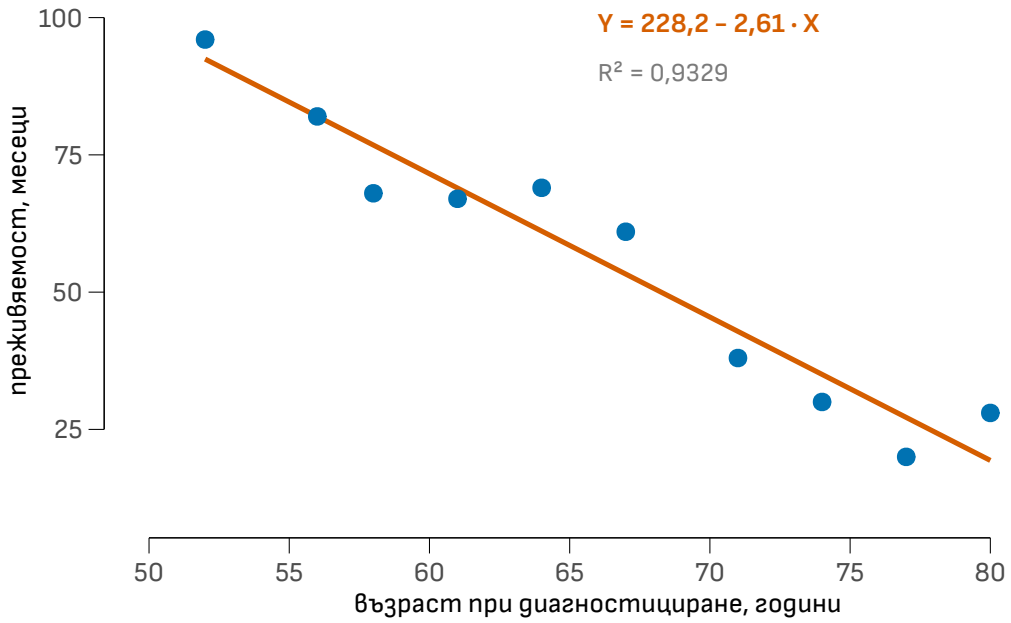
Въпрос 46

Нестандартизираните показатели за заболеваемост в два региона са 15 % за регион А и 19 % за регион Б. Приемайки за стандарт възрастовата структура на регион Б, стандартизираните показатели са 16 % за регион А и 19 % за регион Б. Попълнете липсващите стойности, ако за стандарт се приеме възрастовата структура на **регион А**.

	Нестандартизиран показател	Стандартизиран (стандарт — регион А)	Стандартизиран (стандарт — регион Б)
Регион А	15 %		16 %
Регион Б	19 %		19 %

Въпрос 47

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



Фиг. 8.1: Графика към въпрос 47: връзка между възрастта при диагностициране и преживяемостта при 10 пациенти.

Въпрос 48

Кои са основните рискове при осъществяване на подбор на извадка чрез метода на снежната топка?

Въпрос 49

Проучване сравнява честотата на рецидиви при два вида лечение. Получен е резултат $t = 2{,}080$ при $df = 21$. На колко е равна p-стойността в този случай и какво означава това? Какъв извод може да се направи?

Въпрос 50

Какво представлява явлението автокорелация? Как може да се провери за нейното наличие?

Въпрос 51

Какво представлява биномното разпределение? В какви случаи то може да се приближи с нормалното разпределение?

Въпрос 52

Изследовател иска да сравни измененията в средния индекс на телесна маса преди и след тримесечна кето диета при пациенти със затлъстяване. Индексът на телесна маса е нормално разпределена величина. Какъв тест за проверка на хипотеза следва да бъде приложен?

Въпрос 53

Извадка от 100 новородени е включена в проспективно наблюдение. Средният ръст при раждане е 49 см със стандартно отклонение 1 см. Ръстът е нормално разпределена величина.
Колко от включените новородени имат ръст, по-малък от 49 см или по-голям от 51 см?

Въпрос 54

Какво представляват едностранните тестове за проверка на хипотеза? Кое е тяхното основно предимство и защо?

A Допълнителни понятия за изпита

Приложението събира понятия, които се появяват в изпитните тестове, но не са основни за изчислителните упражнения: избор на графика, модели за подбор на извадка, робастност, биномно разпределение, автокорелация и екстраполация.

A.1 Избор на графика

Видът на графиката се определя от вида на променливата, а не от вкуса на автора.

Таблица A.1: Избор на графика според вида на променливата

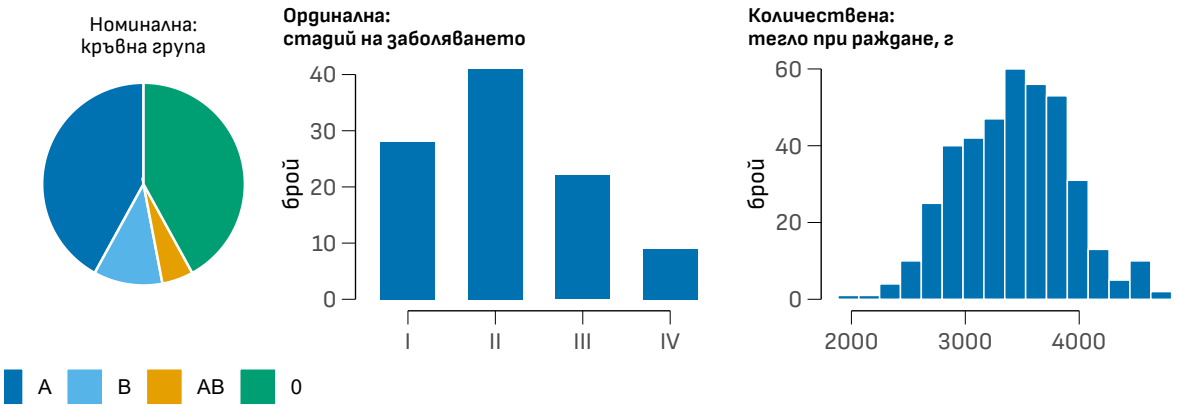
Вид на променливата	Препоръчителна графика	Неподходяща графика
качествена, **номинална**	кръгова диаграма; стълбовидна диаграма	хистограма, боксплот
качествена, **ординална**	**стълбовидна** диаграма (запазва подредбата)	кръгова диаграма
количествена, едно разпределение	хистограма; боксплот	кръгова диаграма
количествена, по групи	боксплот, разделен по групи	кръгова диаграма
две количествени	диаграма на разсейване	стълбовидна диаграма

! Кръгова или стълбовидна

Кръговата диаграма показва **дялове от едно цяло** и затова работи само при номинални категории, при които подредбата няма значение.
При **ординална** променлива подредбата носи информация. Кръговата диаграма я губи, защото окръжността няма начало и край. Затова при ординална скала се предпочита стълбовидна диаграма с категории, подредени по естествения им ред.

⚠ Хистограма и стълбовидна диаграма не са едно и също

Стълбовидната диаграма представя **категории**: стълбовете са отделени с интервали, защото между категориите няма непрекъснатост.
Хистограмата представя **количествена** променлива, групирана в интервали: стълбовете се допират, защото интервалите са съседни части от една непрекъсната скала.



Фиг. A.1: Три графики за три различни вида променливи. Вляво кръгова диаграма за номинална променлива, в средата стълбовидна диаграма за ординална променлива, вдясно хистограма за количествена променлива.

A.2 Модели за подбор на извадка

Начинът, по който е подбрана извадката, определя както възможността за обобщение, така и очаквания размер на грешката.

> ⚠ Двата вида грешка
>
> **Случайна (стохастична) грешка** възниква, защото се наблюдава само част от генералната съвкупност. Тя намалява с увеличаване на обема на извадката и се изразява числено чрез стандартната грешка.
> **Систематична (нестохастична) грешка** възниква от начина на подбор, от неточно измерване или от отпадане на участници. Тя **не** намалява с увеличаване на обема: голяма, но пристрастно подбрана извадка дава прецизна, но невярна оценка.

Вероятностни модели — всяка единица има известна, ненулева вероятност да попадне в извадката. Само при тях случайната грешка може да се изчисли.

Таблица A.2: Вероятностни модели за подбор

Модел	Същност	Очаквана случайна грешка
Проста случайна	всяка единица се тегли независимо	базова
Стратифицирана (типологична)	популацията се дели на еднородни слоеве и се тегли във всеки	**най-малка**
Систематична (механична)	тегли се всеки k-ти елемент от списък	близка до простата случайна
Клъстерна (гнездова)	теглят се цели групи (училища, отделения)	**най-голяма**

Невероятностни модели — вероятността за включване не е известна. При тях случайната

грешка не може да се оцени коректно, а рискът от систематична грешка е висок.

Таблица А.3: Невероятностни модели за подбор

Модел	Същност	Основен риск
Квотна	подбор до запълване на предварително зададени квоти	подборът вътре в квотата не е случаен
По удобство	включват се най-леснодостъпните лица	най-висок риск от систематична грешка
Целенасочена (експертна)	изследователят избира „типични“ случаи	зависи изцяло от преценката му
Снежна топка	участниците насочват към следващи участници	верижен подбор от сходни лица

! Подредба по очаквана случайна грешка

От **най-малка** към **най-голяма**:
стратифицирана (типологична) → проста случайна → квотна → клъстерна → по удобство.
Клъстерната извадка има по-голяма грешка от простата случайна при един и същ обем, защото единиците в един клъстер си приличат и носят по-малко нова информация.

i Метод на снежната топка

Използва се при трудно достижими популации: лица, употребяващи наркотици, пациенти с редки заболявания, мигранти без документи. Няколко начални участници насочват към следващи, те — към още, и извадката „нараства“.
Рисковете са три:

1. **Систематична грешка**, ако началните участници са твърде сходни помежду си по местоживеене, социален статус или нагласи — тогава цялата верига наследява техния профил.
2. **Висока случайна грешка**, особено при малък брой начални участници.
3. **Прекъсване на веригата**: ако участник откаже да участва или не насочи към други, подборът в този клон спира.

А.3 Обем на извадката

Необходимият брой единици на наблюдение зависи от четири неща.

Таблица A.4: От какво зависи необходимият обем на извадката

Ако...	необходимият обем е...
популацията е **еднородна** (малко S)	по-малък
търсеният ефект е **голям**	по-малък
желаното ниво на доверителност е **високо** (99 %)	по-голям
желаната точност е **висока** (тесен интервал)	по-голям

Двата типови изпитни въпроса

Най-малък брой наблюдения е необходим при **еднородна популация и ниско ниво на гаранционна вероятност** (съответно при голям очакван ефект).
Най-голям брой наблюдения е необходим при **разнородна популация и високо ниво на гаранционна вероятност** (съответно при малък очакван ефект).

A.4 Робастност на параметричните тестове

Робастност е свойството на един тест да дава надеждни резултати дори когато предпоставките му са частично нарушени.

Кога параметричните тестове са робастни

t-тестът **запазва** робастност, когато:

- извадките са достатъчно големи — централната гранична теорема осигурява приблизително нормално разпределение на извадковите средни;
- двете извадки са приблизително **еднакви по обем** и по **разсейване**;
- няма изразени екстремни стойности.

t-тестът **губи** робастност, когато:

- извадките са **малки по обем**;
- разпределението е силно асиметрично или има аутлайъри;
- обемите и дисперсиите на двете групи се различават силно.

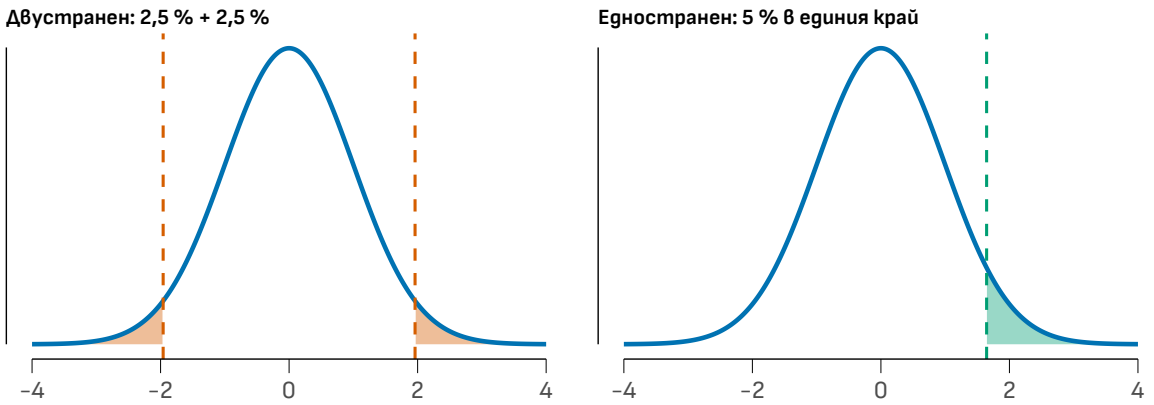
При загуба на робастност се използва непараметричен тест: Mann-Whitney при независими групи и Wilcoxon при свързани.

A.5 Едностранни тестове

Разликата между двустранна и едностранна алтернатива е въведена в Секция 4. Тук е обобщена причината, поради която едностранният тест има по-голяма мощност.

При двустранен тест нивото на значимост $\alpha = 0{,}05$ се разпределя по 2,5 % в двата края на разпределението. При едностранен тест **цялата** критична област от 5 % се съсредоточава в единия край. Затова критичната стойност е по-ниска: 1,645 вместо 1,960 при нормално разпределение.

По-ниският праг означава, че реален ефект **в очакваната посока** се открива по-лесно, тоест тестът има по-висока мощност.



Фиг. A.2: Критични области при двустранен и едностранен тест на едно и също ниво на значимост. При двустранния тест 5 % са разпределени по 2,5 % в двата края; при едностранния цялата критична област е в единия край, което понижава критичната стойност от 1,960 на 1,645.

⚠ Цената на едностранния тест

Едностранният тест **не може** да открие ефект в противоположната посока: ако новото лечение се окаже по-лошо, тестът просто не отхвърля нулевата хипотеза. Затова посоката се обявява **преди** да се видят данните. Смяната на двустранен с едностранен тест след като резултатът е видян, увеличава действителния риск от грешка от I род и е недопустима.

A.6 Биномно разпределение

Когато изходът е **дихотомен** — „да / не", „има / няма", „оздравял / не", — броят на случаите с интересуващия ни изход се описва с **биномно разпределение**.

⚠ Определение

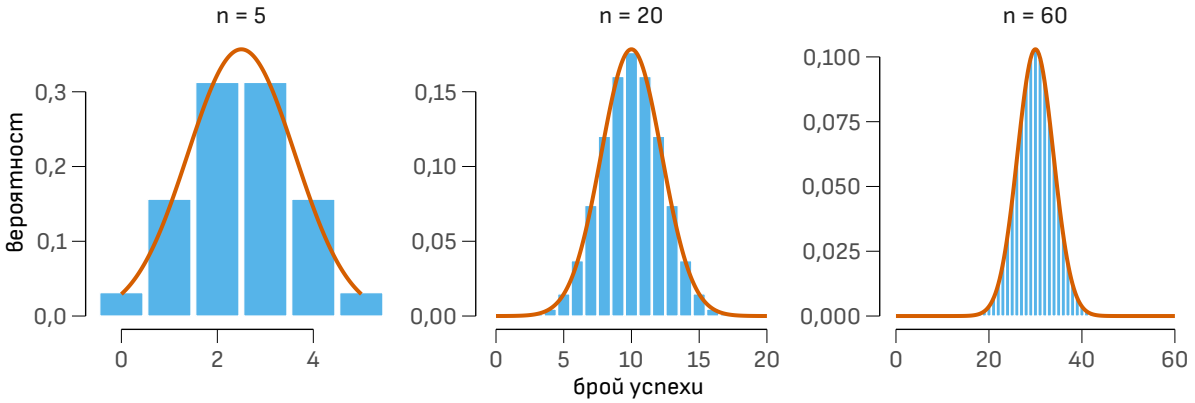
Биномното разпределение описва вероятността да се наблюдават точно k успеха в n независими опита, при които вероятността за успех при всеки опит е една и съща и равна на p.

Изискванията са три: фиксиран брой опити n, независими опити и постоянна вероятност p.

! Кога биномното се приближава с нормално

При **достатъчно голяма извадка** и вероятност p, **не твърде близка до 0 или 1**, биномното разпределение става почти симетрично и се приближава добре с нормалното. Практическото правило е и двата броя — на случаите със събитието и на случаите без него — да бъдат поне 5, тоест $np \geq 5$ и $n(1-p) \geq 5$. Приближението е най-точно при p около 0,5.

Именно това приближение позволява доверителният интервал за относителен дял (Секция 3) и Z-тестът за два дяла (Секция 4) да се строят с критични стойности от нормалното разпределение. При редки събития или малки извадки приближението се разваля и се използват точни методи.



Фиг. A.3: Биномно разпределение при p = 0,5 и три различни обема. С нарастване на n формата става все по-симетрична и все по-близка до нормалната крива, показана с непрекъсната линия.

A.7 Автокорелация

Регресионният модел (Секция 7) предполага, че остатъците са **независими** един от друг.

⚠ Определение

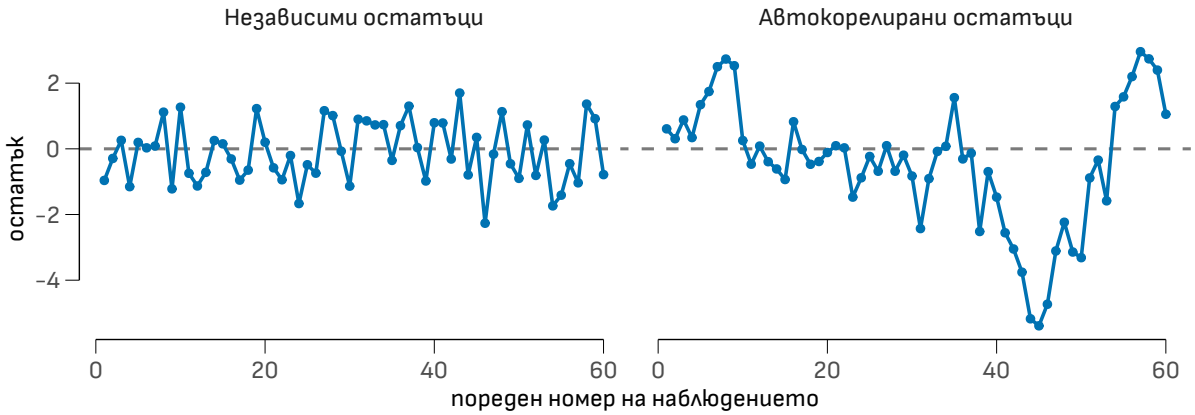
Автокорелацията е явление, при което остатъците не са независими, а са свързани със стойностите си от предходни наблюдения: положителен остатък е последван от положителен, отрицателен — от отрицателен.

Тя се среща най-често при **времеви редове** — месечна заболеваемост, дневен брой хоспитализации, — но и при пространствени данни или при систематично подредени наблюдения.

Проблемът не е в оценките на коефициентите, а в тяхната **точност**: при автокорелация стандартните грешки се подценяват, доверителните интервали изглеждат по-тесни, отколкото са, а p-стойностите — по-малки. Резултатът е привидно значим ефект.

Как се проверява

Най-простата проверка е **графична**: остатъците се нанасят подредени по време или по реда на наблюденията. Ако около нулевата линия се виждат дълги вълни от последователни положителни и отрицателни остатъци вместо случайно разпръскване, има автокорелация. Съществуват и формални тестове, например този на Durbin-Watson.



Фиг. А.4: Остатъци, подредени по време. Вляво разпръскването е случайно и не се откроява модел; вдясно се виждат дълги вълни от последователни положителни и отрицателни остатъци — признак за автокорелация.

А.8 Интерполация и екстраполация

Определения

Интерполация е прогнозиране **вътре** в областта от вече известни стойности, послужили за построяване на модела.
Екстраполация е прогнозиране **извън** тази област.

Интерполацията е обичайното и обосновано приложение на модела. Екстраполацията изисква допълнително и непроверимо допускане: че установената зависимост се запазва и извън наблюдавания диапазон.

Често срещано погрешно твърдение

Екстраполацията **не** повишава валидността на оценката, а я понижава. Колкото по-далеч от наблюдавания диапазон се прогнозира, толкова по-несигурен е резултатът, дори когато формулата продължава да дава числа.

Б Формули

Приложението събира всички формули от курса. До всяка е посочена главата, в която тя е въведена и обяснена.

Относителни величини и стандартизация

Екстензивен (структурен) показател — Секция 1

$$\text{екст.} = \frac{\text{част от явлението}}{\text{цялото явление}} \times 100$$

Интензивен (честотен) показател — Секция 1

$$\text{инт.} = \frac{\text{брой случаи на явлението}}{\text{обем на средата, в която възниква}} \times 100$$

Стандарт (квота по стандарт) — Секция 1

$$S = \frac{\text{брой в подгрупата на стандартната популация}}{\text{общ брой в стандартната популация}} \times 100$$

Стандартизиран показател за подгрупа — Секция 1

$$AIR = \frac{IR \times S}{100}$$

Общият стандартизиран показател е **сборът** на стандартизираните показатели по подгрупи. Общият нестандартизиран показател **не е** сбор на нестандартизираните показатели.

Централна тенденция

Средна аритметична, негрупирани данни — Секция 2

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

Средна аритметична, групирани данни — Секция 2

$$\bar{X} = \frac{\sum (X_u \cdot f)}{\sum f}, \qquad X_u = \frac{\text{долна граница} + \text{горна граница}}{2}$$

Относителен дял — Секция 2

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

Медиана Me — стойността, разделяща възходящо подредения ред на две равни части. **Мода** M_0 — най-често срещаната стойност.

Разсейване

Размах — Секция 2

$$R = X_{max} - X_{min}$$

Стандартно отклонение, негрупирани данни — Секция 2

$$S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$$

Стандартно отклонение, групирани данни — Секция 2

$$S = \sqrt{\frac{\sum(X_u - \bar{X})^2 \cdot f}{n-1}}$$

Интерквартилен размах — Секция 2

$$IQR = Q_3 - Q_1$$

Екстремни са стойностите извън границите $Q_1 - 1{,}5 \cdot IQR$ и $Q_3 + 1{,}5 \cdot IQR$.

Правило на трите сигми (при нормално разпределение) — Секция 2

$$\bar{X} \pm 1S \approx 68\,\%, \qquad \bar{X} \pm 2S \approx 95\,\%, \qquad \bar{X} \pm 3S \approx 99{,}7\,\%$$

Оценка на популационни показатели

Стандартна грешка на средната аритметична — Секция 3

$$SE_{\bar{X}} = S_{\bar{X}} = \frac{S}{\sqrt{n}}$$

Стандартна грешка на относителния дял — Секция 3

$$SE_{\hat{p}} = S_p = \sqrt{\frac{\hat{p}\,(100 - \hat{p})}{n}}$$

Доверителен интервал за средна — Секция 3

$$CI = \bar{X} \pm t_{1-\alpha/2;\,n-1} \cdot SE_{\bar{X}}$$

При голяма извадка се използва приближението $t \approx Z$.

Доверителен интервал за относителен дял — Секция 3

$$CI = \hat{p} \pm Z_{1-\alpha/2} \cdot SE_{\hat{p}}$$

Таблица Б.1: Гаранционни множители

Ниво на доверителност	90 %	95 %	99 %
$Z_{1-\alpha/2}$	1,645	1,960	2,576

Проверка на хипотези

Z-тест за два относителни дяла — Секция 4

$$\hat{p} = \frac{m_1 + m_2}{n_1 + n_2} \times 100$$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\,(100 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

t-тест за две независими извадки — Секция 4

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{SE_{\bar{X}_1}^2 + SE_{\bar{X}_2}^2}}$$

$$df \approx \min(n_1 - 1,\, n_2 - 1)$$

Доверителен интервал за разликата между две средни — Секция 4

$$CI = (\bar{X}_1 - \bar{X}_2) \pm t_{1-\alpha/2;\,df} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Мощност — Секция 4

$$\text{мощност} = 1 - \beta$$

Връзка между качествени признаци

Очаквани честоти — Секция 5

$$E_{ij} = \frac{R_i \times C_j}{N}$$

χ^2-статистика — Секция 5

$$\chi^2 = \sum \frac{(O - E)^2}{E}, \qquad df = (R - 1)(C - 1)$$

Коефициент V на Cramér — Секция 5

$$V = \sqrt{\frac{\chi^2}{N \cdot (k - 1)}}, \qquad k = \min(R,\, C)$$

Коефициент φ при таблица 2 × 2 — Секция 5

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

При таблица 2 × 2 е изпълнено $V = |\varphi|$ и $\chi^2 = Z^2$.

Връзка между количествени признаци

Коефициент на корелация на Pearson — Секция 6

$$r = \frac{\sum (d_x \cdot d_y)}{\sqrt{\sum d_x^2 \cdot \sum d_y^2}}, \qquad d_x = X - \bar{X}, \quad d_y = Y - \bar{Y}$$

Проверка на значимостта на r — Секция 6

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \qquad df = n-2$$

Регресионен коефициент — Секция 7

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

Свободен член — Секция 7

$$a = \bar{Y} - (b \cdot \bar{X})$$

Регресионно уравнение и прогноза — Секция 7

$$\hat{Y} = b \cdot X + a$$

Коефициент на детерминация — Секция 7

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

Знакът пред r съвпада със знака на регресионния коефициент b.

B Статистически таблици

Таблиците дават **критичните стойности** на съответните разпределения при двустранна алтернатива. Начинът на ползване е един и същ за всички:

> **!** Как се чете статистическа таблица
>
> 1. Изчислете тестовата статистика и степените на свобода.
> 2. Намерете реда, съответстващ на вашите степени на свобода.
> 3. Сравнете **абсолютната стойност** на статистиката с числата в реда, като започнете отляво надясно.
> 4. Ако статистиката е **по-малка** от най-лявата стойност, тогава $p > 0{,}10$.
> 5. Ако е между две съседни колони, p е между съответните им нива.
> 6. Ако е **по-голяма** от най-дясната стойност, p е по-малко от нивото на тази колона.
>
> Ако точната стойност на df липсва в таблицата, използва се най-близкият **по-малък** ред — така изводът остава на сигурната страна.

Критични стойности на t-разпределението

Таблица B.1: Критични стойности на t-разпределението при двустранна алтернатива

df	$p = 0{,}10$	$p = 0{,}05$	$p = 0{,}01$
1	6,314	12,706	63,657
2	2,920	4,303	9,925
3	2,353	3,182	5,841
4	2,132	2,776	4,604
5	2,015	2,571	4,032
6	1,943	2,447	3,707
7	1,895	2,365	3,499
8	1,860	2,306	3,355
9	1,833	2,262	3,250
10	1,812	2,228	3,169
11	1,796	2,201	3,106
12	1,782	2,179	3,055
13	1,771	2,160	3,012
14	1,761	2,145	2,977
15	1,753	2,131	2,947
16	1,746	2,120	2,921
17	1,740	2,110	2,898
18	1,734	2,101	2,878

df	$p = 0{,}10$	$p = 0{,}05$	$p = 0{,}01$
19	1,729	2,093	2,861
20	1,725	2,086	2,845
21	1,721	2,080	2,831
22	1,717	2,074	2,819
23	1,714	2,069	2,807
24	1,711	2,064	2,797
25	1,708	2,060	2,787
26	1,706	2,056	2,779
27	1,703	2,052	2,771
28	1,701	2,048	2,763
29	1,699	2,045	2,756
30	1,697	2,042	2,750
35	1,690	2,030	2,724
40	1,684	2,021	2,704
45	1,679	2,014	2,690
50	1,676	2,009	2,678
60	1,671	2,000	2,660
80	1,664	1,990	2,639
100	1,660	1,984	2,626
120	1,658	1,980	2,617
над 120	1,645	1,960	2,576

Последният ред е таблицата на Z

При много големи извадки t-разпределението съвпада с нормалното. Затова редът „над 120“ дава именно критичните стойности на Z: **1,645** при $p = 0{,}10$, **1,960** при $p = 0{,}05$ и **2,576** при $p = 0{,}01$. Тези три числа се използват и при Z-теста за два относителни дяла, и при доверителните интервали.

Критични стойности на хи-квадрат разпределението

Таблица B.2: Критични стойности на χ^2-разпределението

df	$p = 0{,}10$	$p = 0{,}05$	$p = 0{,}01$	$p = 0{,}001$
1	2,706	3,841	6,635	10,828
2	4,605	5,991	9,210	13,816
3	6,251	7,815	11,345	16,266
4	7,779	9,488	13,277	18,467
5	9,236	11,070	15,086	20,515
6	10,645	12,592	16,812	22,458
7	12,017	14,067	18,475	24,322
8	13,362	15,507	20,090	26,124
9	14,684	16,919	21,666	27,877
10	15,987	18,307	23,209	29,588
11	17,275	19,675	24,725	31,264
12	18,549	21,026	26,217	32,909
13	19,812	22,362	27,688	34,528
14	21,064	23,685	29,141	36,123
15	22,307	24,996	30,578	37,697
20	28,412	31,410	37,566	45,315
25	34,382	37,652	44,314	52,620
30	40,256	43,773	50,892	59,703

Най-често използваните редове

При таблица 2 × 2 се използва $df = 1$, при таблица 2 × 3 или 3 × 2 – $df = 2$, при таблица 3 × 3 – $df = 4$, при таблица 4 × 2 – $df = 3$.

Критични стойности на коефициента на корелация

Таблицата позволява значимостта на r да се провери направо, без да се изчислява t. Ако абсолютната стойност на изчисления коефициент е по-голяма от табличната, връзката е значима на съответното ниво.

Таблица B.3: Критични стойности на коефициента на корелация на Pearson

n	$df = n - 2$	$p = 0{,}10$	$p = 0{,}05$	$p = 0{,}01$
5	3	0,805	0,878	0,959
6	4	0,729	0,811	0,917
7	5	0,669	0,754	0,875
8	6	0,621	0,707	0,834
9	7	0,582	0,666	0,798
10	8	0,549	0,632	0,765

n	$df = n - 2$	$p = 0{,}10$	$p = 0{,}05$	$p = 0{,}01$
12	10	0,497	0,576	0,708
15	13	0,441	0,514	0,641
20	18	0,378	0,444	0,561
25	23	0,337	0,396	0,505
30	28	0,306	0,361	0,463
40	38	0,264	0,312	0,403
50	48	0,235	0,279	0,361
60	58	0,214	0,254	0,330
80	78	0,185	0,220	0,286
100	98	0,165	0,197	0,256

⚠ Защо тази таблица е поучителна

Прочетете колоната за $p = 0{,}05$ отгоре надолу. При $n = 8$ е необходимо $|r| \geq 0{,}71$, за да бъде връзката значима, докато при $n = 100$ е достатъчно $|r| \geq 0{,}20$. Следователно **едно и също** число r може да бъде значимо в едно проучване и незначимо в друго. Значимостта зависи от обема на извадката, а силата на връзката — не. Затова винаги се съобщават и трите: r, n и p.

Г Отговори

Отговорите са дадени кратко, за да служат за проверка след самостоятелно решаване. При изчислителните задачи на колоквиума се оценяват формулата, междинните стъпки, мерната единица и словесният извод, а не само крайното число.

Г.1 Към глава 1

Задача 1.1. Възрастовата структура е 23,33 %; 13,67 %; 34,00 % и 29,00 %. Показателите са екстензивни, а сборът им е 100 %.

Задача 1.2. Нестандартизираните показатели за плодовитост са 2,06 ‰ в район А и 1,96 ‰ в район Б. При структурата на район Б като стандарт стандартизираният показател за район А е 1,96 ‰, а този за район Б остава 1,96 ‰. След уеднаквяване на възрастовата структура практически не се наблюдава разлика между районите; незакръглените стойности са 1,959 ‰ и 1,962 ‰.

Задача 1.3. Нестандартизираните показатели за леталитет са 10,50 % в болница А и 9,11 % в болница Б. При структурата на болница Б като стандарт стандартизираните стойности са съответно 9,36 % и 9,11 %. Разликата намалява след стандартизацията, но показателят в болница А остава малко по-висок.

Задача 1.4. При структурата на болница А като стандарт сравнението е 10 ‰ срещу 8 ‰, тоест стандартизираният показател е по-висок в болница А. Новите стойности при структурата на болница Б не могат да се изчислят само от дадените общи показатели. Необходими са показателите по подгрупи и структурата на болница Б. Без тях не може да се гарантира и че посоката на сравнението ще остане същата.

Задача 1.5. Не. Япония има по-възрастна структура на населението, а възрастта е силно свързана със смъртността. Сравнението на общите нестандартизирани показатели смесва влиянието на възрастовата структура с останалите различия между страните. Необходимо е възрастово стандартизирано сравнение.

Задача 1.6. Структурата по отделения е 40,0 % кардиологично, 35,0 % хирургично и 25,0 % неврологично; това са екстензивни показатели, защото знаменателят е цялото от 240 пациенти, разделено на категории. Болничният леталитет е $18/240 \times 100 = 7{,}5\ \%$; това е интензивен показател, защото знаменателят представя пациентите, сред които може да настъпи изходът „смърт“.

Кратък тест: 1 — б; 2 — в; 3 — в; 4 — б; 5 — в.

Г.2 Към глава 2

Задача 2.1. Средите на интервалите са 50, 55, 60, 65, 70, 75 и 80 уд./мин. Претегленият сбор е 2 940, следователно $\bar{X} = 2\,940/45 = 65{,}33$ уд./мин. Извадковото стандартно отклонение е $S = 8{,}88$ уд./мин.

Задача 2.2. Стандартното отклонение е нула само когато всички наблюдения имат една и съща стойност. То не може да бъде отрицателно, защото се получава като квадратен корен от неотрицателна величина.

Задача 2.3. Всички отклонения не могат да бъдат положителни, защото сборът на $X_i - \bar{X}$ винаги е нула. Всички отклонения са нула само когато всички наблюдения са равни на средната.

Задача 2.4. Стойността 4 300 г е на две стандартни отклонения над средната, затова над нея са приблизително 2,5 %, или 50 новородени. Стойността 3 100 г е на едно стандартно отклонение под средната, затова под нея са приблизително 16 %, или 320 новородени. Двете области не се припокриват: общо около 370 новородени.

Задача 2.5. а) категориален, ординална скала; б) количествен дискретен, пропорционална скала; в) категориален, номинална скала; г) количествен непрекъснат, интервална скала; д) количествен дискретен резултат, обичайно третиран като интервална скала; е) категориален дихотомен, номинална скала.

Задача 2.6. $IQR = Q_3 - Q_1 = 50 - 38 = 12$ месеца. Разпределението е ляво изтеглено. Стойността 0 месеца е екстремна; долният мустак достига до 34 месеца.

Задача 2.7. $\bar{X} = 5{,}78$ mg/L, $Me = 4$ mg/L, $Q_1 = 3$ mg/L, $Q_3 = 6$ mg/L и $IQR = 3$ mg/L. Стойността 18 mg/L изтегля средната надясно, затова медианата и интерквартилният размах описват по-подходящо данните.

Задача 2.8. И двете групи имат средна 100. Стандартните отклонения са 1,58 в група А и 15,81 в група Б. Група А е по-хомогенна, защото стойностите ѝ са значително по-слабо разпръснати около една и съща средна.

Кратък тест: 1 — а; 2 — в; 3 — б; 4 — в; 5 — б; 6 — в.

Г.3 Към глава 3

Задача 3.1. Средите на интервалите са 8, 23, 38, 53 и 68 дни. $\bar{X} = 26{,}96$ дни, $S = 12{,}12$ дни и $SE_{\bar{X}} = 12{,}12/\sqrt{250} = 0{,}77$ дни. При $t_{0{,}975;249} = 1{,}970$ получаваме 95 % CI $[25{,}45;\ 28{,}47]$ дни.

Задача 3.2. Средите на интервалите са 3, 8, 13, 18 и 23 месеца. $\bar{X} = 13{,}08$ месеца, $S = 5{,}89$ месеца и $SE_{\bar{X}} = 0{,}34$ месеца. При $t_{0{,}975;299} = 1{,}968$ получаваме 95 % CI $[12{,}41;\ 13{,}75]$ месеца.

Задача 3.3. Точковата оценка остава 56 %. При $Z = 2{,}576$ и $SE_{\hat{p}} = 4{,}96$ процентни пункта 99 % CI е $56 \pm 2{,}576 \cdot 4{,}96 = [43{,}21;\ 68{,}79]$ %. Интервалът е по-широк от 95 % CI, защото по-високото покритие изисква по-голяма критична стойност.

Задача 3.4. Верен отговор: г. Нивото 95 % се отнася до дългосрочното покритие на метода, а не до вероятност за конкретния вече изчислен интервал или за извадковия дял.

Задача 3.5. Обемът нараства четири пъти, затова стандартната грешка намалява $\sqrt{4} = 2$ пъти. При приблизително една и съща критична стойност и ширината на интервала намалява приблизително два пъти.

Задача 3.6. Има 2 наблюдения с подобрение и 18 без подобрение. Условието двата броя да са поне 5 не е изпълнено, затова простият нормален интервал не е подходящ. Предпочита се интервал на Уилсън или точен биномиален интервал.

Задача 3.7. Интервалите се препокриват между 34 % и 38 %. Това не е достатъчно нито за доказване, нито за отхвърляне на разлика. Необходим е доверителен интервал за разликата между дяловете или подходящ тест.

Кратък тест: 1 — в; 2 — в; 3 — б; 4 — в; 5 — а; 6 — в.

Г.4 Към глава 4

Задача 4.1. Дяловете с рецидив са 12,5 % и 22,5 %, а обединеният дял е 17,5 %. Получаваме $Z = -2{,}63$ и двустранно $p = 0{,}0085$. Отхвърляме H_0. При експерименталната терапия извадковият дял с рецидив е с 10 процентни пункта по-нисък; 95 % CI за разликата е $[-17{,}38;\ -2{,}62]$ процентни пункта.

Задача 4.2. Дяловете с едногодишна преживяемост са 83,33 % и 63,33 %. $Z = 5{,}54$ и $p < 0{,}001$, затова отхвърляме H_0. Оценената разлика е 20 процентни пункта, 95 % CI $[13{,}11;\ 26{,}89]$ процентни пункта.

Задача 4.3. Стандартната грешка на разликата е $\sqrt{6^2 + 7^2} = 9{,}22$ mg/dL и $|t| = 15/9{,}22 = 1{,}63$. Без обемите или степените на свобода точната p-стойност от t-разпределението не може да се определи. Нормалното приближение дава $p \approx 0{,}104$ и не води до отхвърляне на H_0 при $\alpha = 0{,}05$.

Задача 4.4. Верен отговор: а. Групите са независими, резултатът не е нормално разпределен и се сравняват три групи, затова е подходящ тестът на Kruskal–Wallis.

Задача 4.5. Верен отговор: б. При непроменени обем, ефект и тест по-ниското α намалява риска от грешка от I род, но увеличава риска от грешка от II род и намалява мощността.

Задача 4.6. Верен отговор: г. Подходящата двустранна алтернатива е, че популационните средни преживяемости са различни. Понеже $p = 0{,}08 > 0{,}05$, не отхвърляме H_0, но не доказваме равенство.

Задача 4.7. При $t = 2{,}96$ и $df = 40$ двустранното $p \approx 0{,}0052 < 0{,}01$. Ако H_0 и предпоставките са верни, вероятността за $|t| \geq 2{,}96$ е около 0,52 %. Отхвърляме H_0 при $\alpha = 0{,}05$ и заключаваме, че има статистически значима разлика във физическата активност между пушачи и непушачи. Посоката и големината изискват описателните оценки.

Задача 4.8. Най-голяма мощност се получава при голям действителен ефект, ниска вариабилност, голяма извадка и по-високо α при равни други условия. На практика α се задава

предварително според допустимия риск от грешка от I род и не се увеличава само за получаване на значим резултат.

Задача 4.9. Подходящ е t-тестът за свързани извадки. $H_0 : \mu_{\text{след}-\text{преди}} = 0$, тоест средната вътреиндивидуална промяна в популацията е нула.

Задача 4.10. Резултатът е статистически значим: интервалът не включва нула и $p < 0{,}001$. Оценяваният ефект е 1,8 mmHg, а интервалът [1,2; 2,4] mmHg показва неговата прецизност. Клиничната значимост не се определя от p; тя изисква сравнение с предварително зададена клинично важна разлика и оценка на ползите и вредите.

Кратък тест: 1 — б; 2 — в; 3 — г; 4 — а; 5 — б; 6 — б.

Г.5 Към глава 5

Задача 5.1. Очакваните честоти са 13,2; 23,4; 23,4 за класическия метод и 8,8; 15,6; 15,6 за аспирационния. Приносите по клетки са 0,776; 0,289; 0,015; 1,164; 0,433 и 0,023, откъдето $\chi^2 = 2{,}70$ при $df = (2-1)(3-1) = 2$. Критичната стойност при $p = 0{,}10$ е 4,605, следователно $p > 0{,}10$ и не отхвърляме H_0. $V = \sqrt{2{,}70/(100 \cdot 1)} = 0{,}16$. Данните не дават достатъчно доказателства за зависимост между метода за аборт и вида на усложнението; извадката е малка и липсата на значимост не доказва липса на връзка.

Задача 5.2. Очакваните честоти са 12,33 и 37,67 за първата и четвъртата група и 24,67 и 75,33 за втората и третата. $\chi^2 = 39{,}21$ при $df = (4-1)(2-1) = 3$; критичната стойност при $p = 0{,}01$ е 11,345, следователно $p < 0{,}01$, а софтуерът дава $p < 0{,}001$. Отхвърляме H_0. $V = \sqrt{39{,}21/(300 \cdot 1)} = 0{,}36$ — силна връзка при $k - 1 = 1$. Най-голям принос има клетката „над 15 г. с оплаквания" — 19,90 от общо 39,21: наблюдават се 28 случая при очаквани 12,33. Има статистически значима, силна зависимост между продължителността на стажа и наличието на неврологични оплаквания.

Задача 5.3. $\chi^2 = 2{,}79$ при $df = 3$; критичната стойност при $p = 0{,}10$ е 6,251, следователно $p > 0{,}10$ и не отхвърляме H_0; $V = 0{,}10$. При зрителните оплаквания дяловете нарастват слабо със стажа (12,9 %; 11,5 %; 13,3 %; 20,0 %), докато при неврологичните нарастването е рязко (8,0 %; 15,0 %; 27,0 %; 56,0 %). Един и същ фактор може да е свързан с един изход и да не е свързан с друг.

Задача 5.4. Очакваните честоти са 30 и 120 за ранната реперфузия и 20 и 80 за късната. $\chi^2 = 15{,}0$ при $df = 1$; критичната стойност при $p = 0{,}01$ е 6,635, следователно $p < 0{,}01$, а софтуерът дава $p < 0{,}001$. $\varphi = -0{,}24$. Отрицателният знак показва, че ранната реперфузия се свързва с **по-ниска** честота на сърдечна недостатъчност: 12,0 % срещу 32,0 %. Има статистически значима, умерена по сила зависимост.

Задача 5.5. а) $df = (4-1)(3-1) = 6$. б) Критичната стойност при $df = 6$ и $p = 0{,}05$ е 12,592; понеже $18{,}0 > 12{,}592$, резултатът е значим. в) $k = \min(4,\ 3) = 3$, тоест $k - 1 = 2$, откъдето $V = \sqrt{18/(400 \cdot 2)} = 0{,}15$. При $k - 1 = 2$ това е слаба до умерена връзка: значима, но не силна.

Задача 5.6. Не. Условието изисква очакваните честоти при таблица 2 × 2 да са достатъчно големи, а 3,2 е под 5. Подходящият избор е точният тест на Fisher.

Задача 5.7. Резултатът е статистически значим, но не и клинично значим. При $V = 0{,}011$ връзката е практически нулева: значимостта се дължи изцяло на огромния обем на извадката, защото χ^2 расте пропорционално на N. Точно затова силата на връзката се съобщава винаги заедно с p-стойността.

Задача 5.8. При $\alpha = 0{,}05$ и 20 независими проверки се очаква средно една значима находка дори когато никоя връзка не съществува. Като съобщава само значимите резултати, изследователят представя очакван брой случайни находки като открития. Коректното поведение е да се обявят всички извършени проверки и при нужда да се приложи корекция за множественост.

Кратък тест: 1 — в; 2 — б; 3 — в; 4 — в; 5 — б; 6 — б.

Г.6 Към глава 6

Задача 6.1. $\bar{X} = 74{,}25$ години и $\bar{Y} = 49{,}375$ месеца; $\sum(d_x d_y) = -755{,}75$, $\sum d_x^2 = 43{,}50$ и $\sum d_y^2 = 15\,115{,}88$, откъдето $r = -755{,}75/\sqrt{43{,}50 \cdot 15\,115{,}88} = -0{,}93$. Проверка на значимостта: $t = -6{,}30$ при $df = 6$; критичната стойност при $p = 0{,}01$ е 3,707, следователно $p < 0{,}01$. Съществува силна обратна, статистически значима връзка: по-високата възраст при диагностициране се свързва с по-кратка преживяемост.

Задача 6.2. $\bar{X} = 0{,}4875$ km и $\bar{Y} = 181{,}25$ pg/μl; $r = -0{,}92$ и $t = -5{,}71$ при $df = 6$, следователно $p < 0{,}01$. Силна обратна, статистически значима връзка: по-голямото изминато разстояние се свързва с по-ниско ниво на NT-proBNP.

Задача 6.3. а) Връзката е низходяща и приблизително линейна в наблюдавания диапазон. б) Наблюдение 8 (0,3 km; 300 pg/μl) е най-отдалечено от общата тенденция, но не е изолирано и не определя само по себе си резултата. в) Коефициентът на Pearson е приемлив. Ако наблюдение 8 бъде премахнато, връзката остава силна и обратна, което говори за устойчивост на извода.

Задача 6.4. Всички d_x не могат да бъдат положителни, защото сборът им винаги е нула. Всички d_x са нула само когато всички стойности на X са равни на средната, тоест променливата е константа. Тогава $\sum d_x^2 = 0$ и знаменателят на формулата става нула: коефициентът не е дефиниран. При липса на изменчивост в X не може да се говори за съгласувано изменение с Y.

Задача 6.5. а) Умерена обратна връзка: по-дългият сън се свързва с по-ниско ниво на кортизол. б) При $n = 60$ и $df = 58$ получаваме $t = -3{,}84$; критичната стойност при $p = 0{,}01$ е около 2,66, следователно $p < 0{,}01$ и връзката е значима. в) При $n = 10$ и $df = 8$ същият коефициент дава $t = -1{,}43$, което е под критичните 1,860 при $p = 0{,}10$; връзката не би била статистически значима. Силата на връзката не се променя, но доказателствената тежест зависи от обема.

Задача 6.6. Не. Най-вероятното обяснение е трета променлива — високата външна температура през летните месеци, която едновременно увеличава консумацията на сладолед и броя на къпещите се, а оттам и удавянията. Това е пример за замъгляващ фактор.

Задача 6.7. Допуснат е капанът „стеснен обхват". Сред професионалните баскетболисти ръстът варира в много тесен диапазон, а изкуственото стесняване на обхвата на едната променлива намалява коефициента на корелация. Изводът за възрастните изобщо е неоправдан.

Задача 6.8. При $n = 20$: $t = 0{,}30 \cdot \sqrt{18}/\sqrt{1 - 0{,}09} = 1{,}33$ при $df = 18$; критичната стойност при $p = 0{,}10$ е 1,734, следователно $p > 0{,}10$ и връзката не е значима. При $n = 200$: $t = 4{,}43$ при $df = 198$, следователно $p < 0{,}001$. Силата на връзката е една и съща в двете проучвания; различава се само точността, с която тя е оценена.

Кратък тест: 1 — б; 2 — б; 3 — в; 4 — а; 5 — в; 6 — б.

Г.7 Към глава 7

Задача 7.1. $\sum X = 594$, $\sum Y = 395$, $\sum XY = 28\,573$ и $\sum X^2 = 44\,148$. Числителят е $8 \cdot 28\,573 - 594 \cdot 395 = 228\,584 - 234\,630 = -6\,046$, а знаменателят $8 \cdot 44\,148 - 594^2 = 353\,184 - 352\,836 = 348$. Следователно $b = -6\,046/348 = -17{,}37$ месеца на година и $a = 49{,}375 - (-17{,}37 \cdot 74{,}25) = 1\,339{,}36$. Моделът е $\hat{Y} = -17{,}37 \cdot X + 1\,339{,}36$. За пациент на 72 години: $\hat{Y} = -17{,}37 \cdot 72 + 1\,339{,}36 = 88{,}5$ месеца. При $R^2 = 0{,}869$ моделът обяснява 86,9 % от вариабилността на преживяемостта, а $r = -\sqrt{0{,}869} = -0{,}93$.

Задача 7.2. $\sum X = 3{,}9$, $\sum Y = 1\,450$, $\sum XY = 639$ и $\sum X^2 = 2{,}09$. Оттук $b = -543/1{,}51 = -359{,}60$ и $a = 181{,}25 + 175{,}31 = 356{,}56$; моделът е $\hat{Y} = -359{,}60 \cdot X + 356{,}56$. Всеки допълнителен изминат километър се асоциира със средно 359,6 pg/µl по-ниско ниво на NT-proBNP, тоест всеки допълнителни 100 метра съответстват на около 36 pg/µl. При $R^2 = 0{,}845$ коефициентът на корелация е $r = -0{,}92$.

Задача 7.3. а) $\sum X = 676$, $\sum Y = 507$, $\sum XY = 43\,761$ и $\sum X^2 = 55\,880$; числителят е $437\,610 - 342\,732 = 94\,878$, а знаменателят $558\,800 - 456\,976 = 101\,824$, откъдето $b = 0{,}93$ точки на минута и $a = 50{,}7 - 0{,}93 \cdot 67{,}6 = -12{,}29$. Моделът е $\hat{Y} = 0{,}93 \cdot X - 12{,}29$. б) При 30 минути четене: $\hat{Y} = 0{,}93 \cdot 30 - 12{,}29 = 15{,}7$ точки. в) $r = +\sqrt{0{,}93} = +0{,}96$; знакът е положителен, защото наклонът е положителен. г) Не. Наблюдателният дизайн допуска и обратна причинност: по-тревожните студенти може да четат повече именно защото се тревожат.

Задача 7.4. а) При увеличаване на дневния прием на сол с 1 g очакваното систолно налягане нараства средно с 1,8 mmHg. б) $|r| = \sqrt{0{,}25} = 0{,}50$ и знакът е положителен, защото $b > 0$, тоест $r = +0{,}50$. в) Не. Обяснената част от вариабилността е само 25 %, тоест три четвърти от измененията се дължат на други фактори; за индивидуална прогноза точността е недостатъчна.

Задача 7.5. а) Първият модел предсказва по-добре, защото обяснява 81 % срещу 36 % от вариабилността. б) $|r| = \sqrt{0{,}81} = 0{,}90$ и $|r| = \sqrt{0{,}36} = 0{,}60$. в) Необходим е знакът на регресионния коефициент, тоест наклонът на правата.

Задача 7.6. а) При увеличаване на възрастта с 1 година очакваната максимална сърдечна честота намалява средно с 0,7 удара в минута. б) $\hat{Y} = 208 - 0{,}7 \cdot 60 = 166$ удара в минута. в) Формално $a = 208$ означава максимална сърдечна честота при възраст нула години.

Стойността е извън обхвата на данните, затова се тълкува само като математическа отправна точка.

Задача 7.7. Прогнозата при 15-годишно дете е **екстраполация**: моделът е построен върху възрасти 40–60 години и няма основание да се приема, че същата права важи извън този диапазон. Формулата ще даде число, но то не е обосновано.

Задача 7.8. Вторият модел. Високият R^2 при 5 наблюдения е слабо доказателство: през малко на брой точки права може да се прекара почти идеално, без това да означава, че връзката ще се потвърди при нови данни. Моделът върху 5 000 наблюдения дава по-нисък, но много по-надеждно оценен R^2.

Кратък тест: 1 — б; 2 — в; 3 — б; 4 — в; 5 — б; 6 — в.

Г.8 Към колоквиума

Пълните решения на четирите варианта са дадени в Секция 8. По-долу са отговорите на трите части на теста за подготовка и на отворените въпроси.

Тест за подготовка, част 1

1 — б. Знаменателят е цялото (всички преминали пациенти), разделено на категории; сборът на дяловете е 100 %.

2 — в. Групата, чиято структура е приета за стандарт, запазва общия си показател. Това е удобна проверка на изчисленията.

3 — б. При изтеглено разпределение средната се измества към екстремните стойности; медианата и интерквартилният размах описват данните по-добре.

4 — в. Границите 2 400 г и 4 400 г отстоят на две стандартни отклонения от средната, тоест обхващат приблизително 95 % от наблюденията: $0{,}95 \cdot 400 = 380$.

5 — в. Стандартната грешка описва разсейването на извадковата средна и намалява с $\sqrt{n}$; тя е по-малка от стандартното отклонение.

6 — в. Нивото 95 % се отнася до дългосрочното покритие на **метода**, а не до вероятност за конкретния вече изчислен интервал.

7 — б. По-ниското α намалява риска от грешка от I род, но увеличава риска от грешка от II род, тоест намалява мощността.

8 — б. Едни и същи лица са измерени двукратно, следователно дизайнът е зависим.

9 — в. При $df = 60$ критичната стойност за $p = 0{,}10$ е 1,671. Понеже $1{,}42 < 1{,}671$, следва $p > 0{,}10$ и не отхвърляме H_0.

10 — в. $df = (3-1)(4-1) = 6$.

11 — в. Нормалното разпределение е условие за параметричните тестове за количествени признаци, а не за χ^2-теста, който работи с честоти.

12 — б. χ^2 расте с обема на извадката, докато V е стандартизирана мярка за силата на връзката.

13 — б. Коефициентът на Pearson измерва само линейната част от връзката. При U-образна зависимост той може да е близък до нула въпреки ясната нелинейна връзка.

14 — в. Регресионният коефициент показва средната промяна в Y при увеличаване на X с една единица.

15 — б. $|r| = \sqrt{0{,}64} = 0{,}80$, а знакът се определя от наклона, който е отрицателен.

Тест за подготовка, част 2

16 — а. Кръговата диаграма представя дялове от едно цяло и е подходяща при номинална скала, при която подредбата на категориите няма значение. Стълбовидната също е допустима, но кръговата е препоръчителната при чисто номинален признак.

17 — г. Общият нестандартизиран показател не е нито сума, нито произведение на подгруповите показатели. Той се изчислява наново, от общия брой събития и общия брой наблюдения.

18 — а. Категориите „рак на белия дроб“, „рак на гърдата“, „рак на дебелото черво“ и „друго“ нямат логическа подредба, следователно скалата е номинална.

19 — а. Свободният член показва къде правата пресича оста Y и не носи информация за наклона. Положителен свободен член е съвместим и с положителен, и с отрицателен наклон.

20 — б. Интервалите [21 %; 28 %] и [32 %; 42 %] не се препокриват, тоест разликата е статистически значима. При терапия А честотата на рецидиви е по-ниска, следователно тя е по-ефективна.

21 — а. От най-малка към най-голяма случайна грешка: типологична (стратифицирана) извадка, квотна извадка, клъстерна извадка. Стратификацията намалява разсейването вътре в слоевете, а клъстерната извадка го увеличава, защото единиците в един клъстер си приличат.

22 — а. Правилото е $p < 0{,}05$ за отхвърляне на H_0. Стойността $p = 0{,}050$ не е по-малка от 0,05, затова не отхвърляме нулевата хипотеза и данните не позволяват твърдение за разлика. Това е граничен случай: при $p = 0{,}049$ изводът щеше да бъде обратният, което показва защо решението не бива да се свежда до едно число.

23 — в. Медианата зависи от **позицията** на наблюденията, а не от техните стойности, затова е най-устойчива на екстремни стойности. Средната аритметична, стандартното отклонение и стандартната грешка се изчисляват от самите стойности и се влияят силно.

24 — а. $p = 0{,}006 < 0{,}05$, следователно нулевата хипотеза се отхвърля. Тестът на McNemar не е приложим при независими групи и качествени признаци; използваният t-тест предполага приблизително нормално разпределение; а алтернативната хипотеза твърди, че средните са **различни**.

25 — в. Екстраполацията не повишава, а понижава валидността на оценката, защото изисква допълнителното и непроверимо допускане, че зависимостта се запазва извън наблюдавания диапазон.

26 — а. Робастността се губи при малки извадки. Отсъствието на аутлайъри и сходните обеми и дисперсии са условия, при които тестът **запазва** робастност.

27 — б. При големи извадки t-разпределението се доближава до нормалното; при df над 120 критичните стойности практически съвпадат с 1,645; 1,960 и 2,576.

28 — б. В изследваната извадка честотата е известна точно и е равна на 10 %. Доверителният интервал оценява **популационната** честота, а не тази в извадката.

29 — г. Най-малък обем е необходим при еднородна популация (малко разсейване) и ниско ниво на гаранционна вероятност (по-малка критична стойност).

30 — в. $SE_{\bar{X}} = S/\sqrt{n} = 22/\sqrt{121} = 22/11 = 2$ месеца.

Тест за подготовка, част 3

31 — б. Тестът на McNemar се прилага при **свързани** извадки и дихотомен качествен признак. Тук групите са независими, а признакът е количествен, следователно твърдението е грешно.

32 — в. Критичните стойности от Z-разпределението се използват, когато популационното стандартно отклонение е известно или е оценено надеждно от достатъчно голяма извадка.

33 — а. Най-голям обем се изисква при разнородна популация (голямо разсейване) и високо ниво на гаранционна вероятност.

34 — г. Ръстът има естествена нула и допуска отношения, следователно скалата е пропорционална.

35 — в. Интервалите [21 %; 28 %] и [24 %; 38 %] се препокриват в областта 24–28 %. Само въз основа на препокриването не може да се твърди, че разлика има. Забележете, че препокриването също така **не доказва** равенство; за категоричен извод е нужен доверителен интервал за разликата или подходящ тест.

36 — б. Обратната подредба на въпрос 21: от най-голяма към най-малка случайна грешка — клъстерна, квотна, типологична извадка.

37 — б. $p = 0{,}0051 < 0{,}05$, следователно разликата е статистически значима. По-ниската честота на рецидиви (24 %) е при терапия Б, тоест тя е по-ефективната.

38 — в. Отрицателният регресионен коефициент означава отрицателен наклон: с увеличаване на X очакваната стойност на Y намалява.

39 — г. Ординалната скала подрежда категориите логически, но не гарантира равни разстояния между тях и не е най-точната скала. Всяка единица принадлежи към точно една категория.

40 — а. Стандартното отклонение се изчислява от квадратите на отклоненията и затова реагира най-силно на екстремни стойности. Интерквартилният размах и модата са устойчиви, а нивото на значимост се задава от изследователя.

41 — в. Стандартизираните показатели по подгрупи **се събират** и сборът им дава общия стандартизиран показател.

42 — г. $SE_{\bar{X}} = 9/\sqrt{81} = 9/9 = 1$ месец.

43 — в. Робастността се запазва при приблизително еднакви по обем и по разсейване извадки.

44 — б. Доверителният интервал оценява **популационния относителен дял**, а не вероятността за отделно дете. Интервалът не е твърдение за индивид.

45 — б. При ординална скала подредбата на категориите носи информация. Стълбовидната диаграма я запазва, докато кръговата я губи, защото окръжността няма начало и край.

Отворени въпроси

Въпрос 46. Регион А запазва своя показател от **15 %**, защото неговата структура е приета за стандарт. За регион Б очакваният отговор е **18 %**.

	Нестандартизиран	Стандарт — регион А	Стандарт — регион Б
Регион А	15 %	**15 %**	16 %
Регион Б	19 %	**18 %**	19 %

Изводът е един и същ и при двата стандарта: заболеваемостта в регион Б е по-висока, а разликата е около 3 процентни пункта.

Методологична бележка. Точната стойност за регион Б не следва еднозначно само от общите показатели: за нейното изчисление са необходими показателите по възрастови подгрупи и структурата на регион А. Сигурни са две неща — регион А задължително запазва 15 %, а посоката на разликата се запазва.

Въпрос 47. Моделът е $\hat{Y} = 228{,}2 - 2{,}61 \cdot X$. Регресионният коефициент $b = -2{,}61$ е отрицателен, тоест зависимостта е **обратна**: при увеличаване на възрастта при диагностициране с **1 година** очакваната преживяемост намалява средно с **2,61 месеца**.

Коефициентът на детерминация $R^2 = 0{,}9329$ означава, че **93,3 %** от вариабилността на преживяемостта се обяснява с възрастта при диагностициране; останалите 6,7 % се дължат на други фактори и на случайна вариация.

Коефициентът на корелация е $|r| = \sqrt{0{,}9329} = 0{,}97$, а знакът е отрицателен, защото наклонът на правата е отрицателен, тоест $r \approx -0{,}97$ — много силна обратна корелация.

Въпрос 48. Основните рискове при подбор чрез метода на снежната топка са три. Първо, **систематична грешка**: ако началните участници са твърде сходни по местоживеене, социален статус или нагласи, цялата верига наследява техния профил и извадката не представя популацията. Второ, **висока случайна грешка**, особено когато началните участници са

малко на брой. Трето, **прекъсване на веригата**: при отказ за участие или при участник, който не насочва към други, подборът в този клон спира. Освен това вероятността за включване не е известна, което не позволява коректна оценка на случайната грешка.

Въпрос 49. При $df = 21$ критичната стойност за $p = 0{,}05$ е точно **2,080**. Изчисленото $t = 2{,}080$ съвпада с нея, тоест $p = 0{,}05$.

По правилото $p \geq 0{,}05$ не отхвърляме нулевата хипотеза. Резултатът не е статистически значим при $\alpha = 0{,}05$ и не дава основание да се твърди, че между двата вида лечение има разлика по отношение на честотата на рецидиви. Това не доказва, че разлика няма: стойността е точно на границата и по-голяма извадка би могла да доведе до различен извод.

Въпрос 50. Автокорелацията е явление, при което остатъците от регресионния модел не са независими, а са свързани със стойностите си от предходни наблюдения. Среща се най-често във времеви редове, но и в пространствени или систематично подредени данни.

Последицата е подценяване на стандартните грешки: доверителните интервали изглеждат по-тесни, а p-стойностите — по-малки от действителните, което води до привидно значими резултати.

Проверява се преди всичко **графично**: остатъците се нанасят подредени по време или по реда на наблюденията и се търсят дълги вълни от последователни положителни и отрицателни стойности вместо случайно разпръскване. Съществуват и формални тестове, например този на Durbin–Watson.

Въпрос 51. Биномното разпределение описва вероятността да се наблюдават точно k успеха в n независими опита, при които вероятността за успех при всеки опит е една и съща и равна на p. То служи за моделиране на честотата на дихотомни изходи от типа „да / не“, „има / няма“.

Биномното разпределение може да се приближи с нормално при **достатъчно голяма извадка** и вероятност p, която не е твърде близка до 0 или до 1. Практическо правило е броят на случаите със събитието и броят без него да бъдат поне 5, тоест $np \geq 5$ и $n(1-p) \geq 5$; приближението е най-добро при p около 0,5. Именно на това приближение се основават доверителният интервал за относителен дял и Z-тестът за два дяла.

Въпрос 52. Индексът на телесна маса е количествена, нормално разпределена величина, измерена **двукратно при едни и същи пациенти** — преди и след диетата. Дизайнът е зависим, следователно се прилага **t-тест за две свързани извадки**.

Въпрос 53. Ръстът е нормално разпределен със средна 49 см и стандартно отклонение 1 см.

- Стойността 49 см съвпада със средната, следователно под нея са **50 %** от новородените, тоест 50 души.
- Стойността 51 см е на две стандартни отклонения над средната, следователно над нея са приблизително **2,5 %**, тоест около 2 души.

Двете области не се препокриват и общо около **52** от 100-те новородени имат ръст под 49 см или над 51 см.

Въпрос 54. Едностранните тестове са тестове, при които алтернативната хипотеза има предварително зададена посока: изследователят очаква едната стойност да бъде само по-голяма или само по-малка от другата.

Основното им предимство е **по-голямата статистическа мощност**. Причината е, че цялата критична област α се съсредоточава в единия край на разпределението вместо да се разделя между двата. Това понижава критичната стойност — например от 1,960 на 1,645 при нормално разпределение — и прави теста по-чувствителен към разлики в очакваната посока.

Цената на това предимство е, че ефект в противоположната посока не може да бъде открит. Затова посоката се обявява **преди** да се видят данните; преминаването към едностранен тест след като резултатът е известен увеличава действителния риск от грешка от I род.

Д Речник на основните понятия

Понятията са подредени по азбучен ред. В скоби е посочена главата, в която понятието е въведено.

Абсолютна величина — наименовано число, получено чрез пряко измерване или преброяване. (Секция 1)

Алтернативна хипотеза H_1 — твърдението, противоположно на нулевата хипотеза: наблюдаваната разлика или връзка е съществена. (Секция 4)

Вариационен ред — подредено представяне на данните заедно с честотите им; бива степенен (по отделни стойности) и интервален (по интервали). (Секция 2)

Генерална съвкупност — всички единици, за които се отнася изводът от изследването. (Секция 2)

Грешка от I род α — отхвърляне на вярна нулева хипотеза, тоест обявяване за откритие на нещо, което не съществува. (Секция 4)

Грешка от II род β — неотхвърляне на грешна нулева хипотеза, тоест пропускане на действително съществуваща разлика. (Секция 4)

Гаранционен множител Z — критичната стойност, с която се умножава стандартната грешка при построяване на доверителен интервал. (Секция 3)

Дескриптивна статистика — описание на данните в извадката. (Секция 2)

Доверителен интервал — интервал от стойности, построен така, че при многократно повторение на процедурата да съдържа истинския параметър в предварително зададен дял от случаите. (Секция 3)

Екстензивен показател — структурен показател, който показва как едно цяло се разпределя между категории; сборът е 100 %. (Секция 1)

Екстраполация — прилагане на установена връзка извън обхвата на данните, върху които тя е построена. (Секция 6)

Екстремна стойност (аутлайър) — наблюдение извън границите $Q_1 - 1{,}5 \cdot IQR$ и $Q_3 + 1{,}5 \cdot IQR$. (Секция 2)

Замъгляващ фактор — трета променлива, свързана едновременно с изследвания фактор и с изхода, която изкривява наблюдаваната връзка. (Секция 1)

Извадка — подмножеството от единици, което действително е изследвано. (Секция 2)

Инферентна статистика — пренасяне на изводите от извадката към генералната съвкупност. (Секция 3)

Интензивен показател — честотен показател, който показва колко често възниква събитие в популацията под риск. (Секция 1)

Интерквартилен размах IQR — разликата между третия и първия квартил; мярка за разсейване около медианата. (Секция 2)

Квартили Q_1, Q_2, Q_3 — стойностите, разделящи подредения статистически ред на четири равни части. (Секция 2)

Клинична значимост — оценка на практическото значение на установен ефект, независима от статистическата значимост. (Секция 4)

Коефициент на детерминация R^2 — делът от разсейването на зависимата променлива, обяснен от модела. (Секция 7)

Коефициент на корелация на Pearson r — мярка за силата и посоката на линейната връзка между две количествени променливи; приема стойности от -1 до $+1$. (Секция 6)

Колинеарност — силна корелация между два или повече предиктора в множествен регресионен модел. (Секция 7)

Контингентна таблица (таблица на съчетаемост) — таблица, разпределяща наблюденията едновременно по две качествени променливи. (Секция 5)

Медиана Me — стойността, разделяща възходящо подредения ред на две равни части. (Секция 2)

Мода M_0 — най-често срещаната стойност в статистическия ред. (Секция 2)

Мощност — вероятността да се открие действително съществуваща разлика; равна на $1 - \beta$. (Секция 4)

Нулева хипотеза H_0 — твърдението, че разлика или връзка не съществува и наблюдаваното различие е случайно. (Секция 4)

Ниво на значимост α — предварително избраният допустим риск от грешка от I род, обичайно 0,05. (Секция 4)

Нормално разпределение — симетрично камбановидно разпределение, при което средната, медианата и модата съвпадат. (Секция 2)

Обединен относителен дял — общият дял, изчислен от двете групи заедно и използван в знаменателя на Z-теста. (Секция 4)

Остатък — разликата между наблюдаваната и прогнозираната стойност на зависимата променлива. (Секция 7)

Относителна величина — резултат от деление на две абсолютни величини. (Секция 1)

Относителен дял $\hat{p}$ — делът на наблюденията с дадена характеристика, изразен в проценти. (Секция 2)

Очаквана честота E — честотата, която би се наблюдавала при вярна нулева хипотеза за независимост. (Секция 5)

Параметър — показател на генералната съвкупност; обикновено неизвестен. (Секция 3)

p-стойност — вероятността при вярна нулева хипотеза да се получи такъв или по-екстремен резултат. (Секция 4)

Правило на трите сигми — при нормално разпределение приблизително 68 %, 95 % и 99,7 % от наблюденията попадат съответно в границите $\bar{X} \pm 1S$, $\bar{X} \pm 2S$ и $\bar{X} \pm 3S$. (Секция 2)

Прогнозирана стойност $\hat{Y}$ — средно очакваната стойност на зависимата променлива при дадена стойност на предиктора. (Секция 7)

Размах R — разликата между най-голямата и най-малката стойност. (Секция 2)

Регресионен коефициент b — средната промяна в зависимата променлива при увеличаване на предиктора с една единица. (Секция 7)

Свободен член a — стойността на прогнозираната зависима променлива при $X = 0$; отсечка по оста Y. (Секция 7)

Скала за измерване — номинална, ординална, интервална или пропорционална; определя кои методи са допустими. (Секция 2)

Стандарт — структурата, приета за обща при стандартизация; сборът на квотите е 100 %. (Секция 1)

Стандартизация (пряк метод) — преобразуване на общите коефициенти, с което се отстранява влиянието на различията в състава на сравняваните групи. (Секция 1)

Стандартна грешка SE — мярка за прецизността на извадкова оценка; описва разсейването на оценката от извадка на извадка. (Секция 3)

Стандартно отклонение S — мярка за разсейването на отделните наблюдения около средната аритметична. (Секция 2)

Статистика — показател, изчислен от извадката. (Секция 3)

Статистическа значимост — резултат, при който p-стойността е под избраното ниво α. (Секция 4)

Статистически признак — характеристика, по която се различават единиците на наблюдение. (Секция 2)

Степени на свобода df — параметър на разпределението, който определя кой ред от статистическата таблица се използва. (Секция 4)

Таблица на съчетаемост — вж. *контингентна таблица*.

Тестова статистика — числото, изчислено от данните (Z, t, χ^2), чиято стойност се сравнява с критичните стойности. (Секция 4)

Хи-квадрат χ^2 — тестова статистика за независимост между два качествени признака. (Секция 5)

Цел на теста за независимост — да се провери дали разпределението на едната променлива е едно и също при всички категории на другата. (Секция 5)

Централна тенденция — обобщаваща стойност, представяща наблюденията като цяло: средна аритметична, медиана или мода. (Секция 2)

Коефициент V **на Cramér** — стандартизирана мярка за силата на връзката в контингентна таблица с произволен размер; приема стойности от 0 до 1. (Секция 5)

Коефициент φ — мярка за силата и посоката на връзката в таблица 2 × 2; интерпретира се като корелационен коефициент. (Секция 5)