Дескриптивна статистика. Доверителен интервал

Централна тенденция и разсейване, стандартна грешка, 95 % доверителен интервал за средна и за относителен дял — с решени примери стъпка по стъпка

Упражнение 2 · Медицинска статистика · 2026/2027

гл. ас. д-р Костадин Костадинов, д.м.

Катедра по социална медицина и обществено здраве · Медицински университет – Пловдив

Какво ще можете след това упражнение

1. Да определяте **вида на признака** и **скалата**, на която се измерва.
2. Да изчислявате **средна аритметична** (при негрупирани и групирани данни), **медиана** и **мода**.
3. Да изчислявате **размах**, **стандартно отклонение** и **интерквартилен размах**; да избирате $\overline{X} \pm S$ или $\mathrm{Me}\ \ (\mathrm{IQR})$.
4. Да различавате **стандартно отклонение** от **стандартна грешка**.
5. Да строите **95 % доверителен интервал** за средна и за относителен дял и да го тълкувате правилно.

План на часа (90 мин)

1. Домашна 1: отговори — 5 мин
2. Признаци и скали — 6 мин
3. Централна тенденция — 15 мин
4. Разсейване — 15 мин
5. Доверителен интервал — 22 мин
6. Задача в час — 10 мин
7. Текуща проверка (1 т.) — 15 мин

Калкулатор и лист с формули. Междинните стойности **не** се закръгляват.

Домашна 1: отговори

Задача 1 · Смъртност в две общини

- Обща смъртност: А **9,000 ‰**, Б **18,025 ‰**
- Стандарт — структурата на А (60 / 30 / 10 %): А **9,000 ‰**, Б **7,550 ‰**
- Б е „по-лоша“ само заради по-възрастното си население; във **всяка** възрастова група смъртността в Б е по-ниска.

Задача 2 · Леталитет в две болници

- Нестандартизиран: А **3,125 %**, Б **5,300 %**
- Стандарт Б: А **6,625 %**, Б **5,300 %**
- Стандарт А: А **3,125 %**, Б **2,500 %**
- Стандартизираните стойности зависят от стандарта, но посоката е една и съща: Б е по-добра; изглежда по-лоша, защото лекува повече болни в отделението по интензивно лечение.

СЪДЪРЖАНИЕ

1. **Признаци и скали за измерване**
2. Показатели за централна тенденция
3. Показатели за разсейване
4. От извадката към популацията: доверителен интервал
5. Задача в час и домашна работа

Видове признаци и скали

Скала	Характеристика	Пример
номинална	категории без ред; дихотомна — само две	кръвна група; пол
ординална	категории с ред; разликите не са измерими	стадий I–IV; степен на болка
интервална	равните разлики са равни по смисъл; „0“ не е липса на признака	температура в °C
пропорционална	числа; „0“ е липса; има смисъл и отношението	тегло, ръст, брой дни

Признаци

- **качествени** (категориални): номинални, ординални
- **количествени**: дискретни (брой раждания) и непрекъснати (тегло, кръвно налягане)

Защо има значение

Скалата определя **кои показатели** са допустими: средна и S — само за количествени; за качествени — **относителен дял**.

Бърза проверка: каква е скалата?

а) стадий на бъбречно заболяване (I–V)
б) брой цигари на ден
в) кръвна група
г) телесна температура в °C
д) наличие на алергия (да / не)
е) систолно кръвно налягане в mmHg

Отговори

а) ординална · б) пропорционална, дискретна · в) номинална · г) интервална · д) номинална, дихотомна · е) пропорционална, непрекъсната

СЪДЪРЖАНИЕ

1. Признаци и скали за измерване
2. **Показатели за централна тенденция**
3. Показатели за разсейване
4. От извадката към популацията: доверителен интервал
5. Задача в час и домашна работа

Средна аритметична: негрупирани данни

$$\overline{X} = \frac{\sum X_i}{n}$$

X_i — стойността на i-тото наблюдение; n — броят наблюдения.

Пример: пулс на 10 студенти

90, 80, 89, 72, 73, 74, 78, 86, 84, 82 уд./мин

$\overline{X} = 808/10 =$ **80,8 уд./мин**

Медиана и мода

- **Медиана** Me — делѝ подредения ред на две равни части. При нечетен n — средното наблюдение ($(n+1)/2$-ото); при четен — средната на двете средни стойности: подреден пулс ... 80, **82** ...; $\mathrm{Me} = (80+82)/2 =$ **81 уд./мин**.
- **Мода** M_0 — най-честата стойност; може да няма или да са няколко. В медицината се ползва рядко.

Резултатът винаги се пише **с мерна единица**.

Групиране на данните: вариационни редове

Степенен вариационен ред

Всяка стойност и срещу нея — колко пъти се среща. **Не губим** точност.

Интервален вариационен ред

Стойностите се групират в интервали. **Губим** точност: знаем само интервала.

Две правила

1. Интервалите са с **еднаква ширина** (когато е приложимо; тук винаги).
2. Всяко наблюдение попада в **точно един** интервал: 1–5, 6–10, 11–15 (не 1–5, 5–10).

Примерът на упражнението

Преживяемост (в цели месеци) на $n = 30$ пациенти с експериментална терапия:

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

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

Средна аритметична: интервален ред

$$\overline{X} = \frac{\sum X_u \cdot f}{\sum f}$$

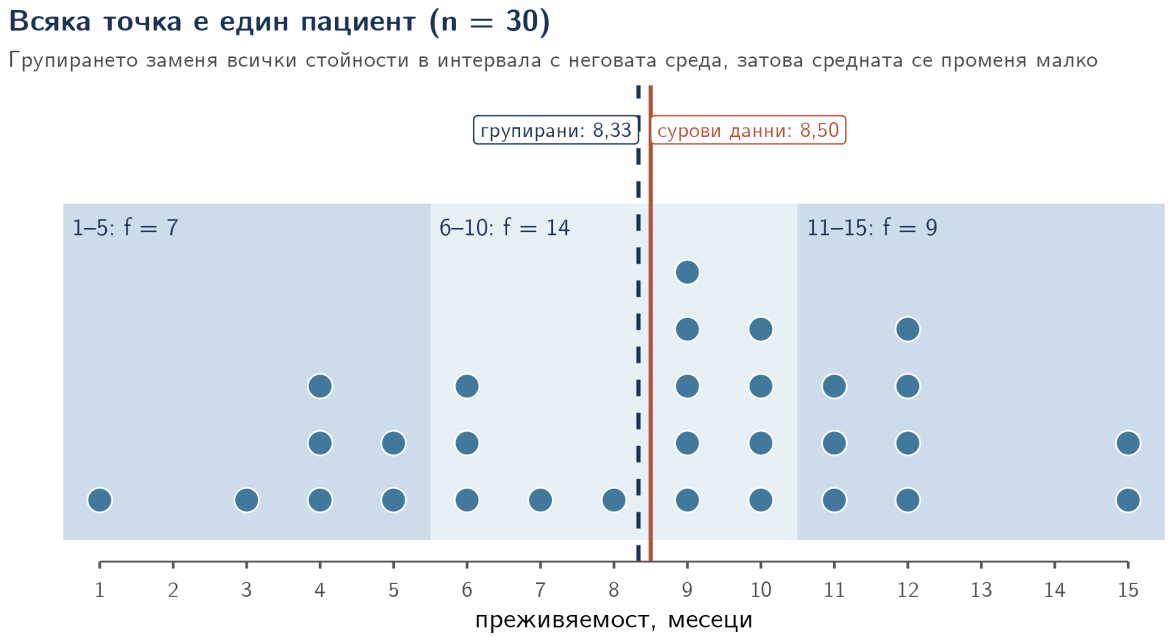
$X_u = \frac{\text{долна} + \text{горна граница}}{2}$ — среда на интервала; f — честота.

В листа с формули: $\overline{X}_w = \sum(x_i \cdot w_i) / \sum w_i$ — същата формула.

Интервал	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**

$\overline{X} = 250/30 = 8{,}333 \approx$ **8,33 месеца. Не закръгляйте до 8 преди следващите стъпки.**

Цената на групирането



Сурови данни: $\overline{X} = 255/30 =$ **8,50 месеца**, $\mathrm{Me} =$ **9 месеца**. Групирани: **8,33 месеца**. Когато суровите данни са налице, изчисляваме от тях.

Качествен признак: относителен дял

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

m — брой наблюдения с признака; n — всички наблюдения.

Пример (пример II на упражнението)

Безглутенова диета: подобрение при **56 от 100** пациенти.

$\hat{p} = 56/100 \times 100 =$ **56 %**

Това е **точкова оценка** на неизвестния популационен дял P. Колко е точна — ще видим с доверителния интервал.

СЪДЪРЖАНИЕ

1. Признаци и скали за измерване
2. Показатели за централна тенденция
3. **Показатели за разсейване**
4. От извадката към популацията: доверителен интервал
5. Задача в час и домашна работа

Стандартно отклонение: негрупирани данни

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

- **квадратираме**, защото $\sum\left(X_i - \overline{X}\right) = 0$ винаги
- $n-1$: оценяваме от **извадка**
- коренът връща **мерната единица**

$S = \sqrt{383{,}6/9} = \sqrt{42{,}62} =$ **6,53 уд./мин**

Запис: $80{,}8 \pm 6{,}53$ **уд./мин**

Размах: $R = 90 - 72 = 18$ уд./мин — зависи само от две стойности.

X_i	$X_i - \overline{X}$	$\left(X_i - \overline{X}\right)^2$
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
808	**0,0**	**383,60**

Стандартно отклонение: интервален ред

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

$S = \sqrt{396{,}67/29} = \sqrt{13{,}68} = \mathbf{3{,}70}$ **месеца**

Запис: $8{,}33 \pm 3{,}70$ **месеца**

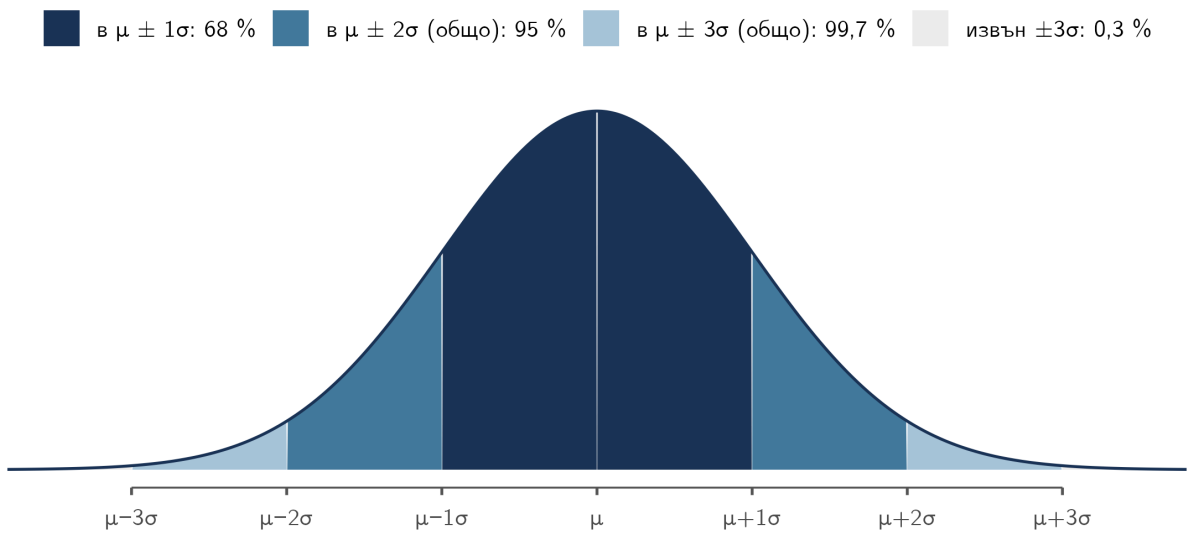
От суровите данни: $S = 3{,}50$ месеца.

Интервал	f	X_u	$X_u - \overline{X}$	$\left(X_u - \overline{X}\right)^2$	$\left(X_u - \overline{X}\right)^2 \cdot f$
1–5	7	3	–5,333	28,444	199,111
6–10	14	8	–0,333	0,111	1,556
11–15	9	13	4,667	21,778	196,000
Общо	**30**				**396,667**

Правило на трите сигми

Само при нормално разпределение

Пример: тегло 80 ± 10 кг → около 95 % от хората са между 60 и 100 кг



- $\mu \pm 1\sigma$ — около **68 %** от наблюденията
- $\mu \pm 2\sigma$ — около **95 %**
- $\mu \pm 3\sigma$ — около **99,7 %**

Процентите са натрупани. В извадка заместваме с $\overline{X} \pm kS$ — приблизително.

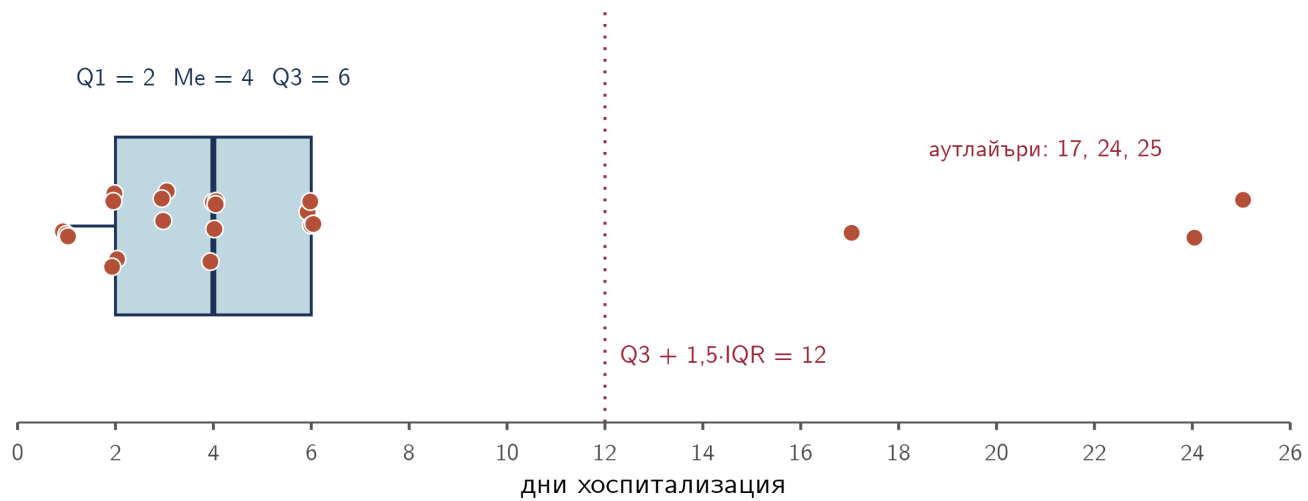
Внимание

Описва **отделните наблюдения** и важи **само при нормално разпределение**. Не е доверителен интервал.

Изтеглено разпределение: медиана и IQR

Дясно изтеглено разпределение (n = 22)

Средна 5,91 дни, медиана 4 дни: средната е издърпана от трите дълги престоя



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

Запис: $\mathrm{Me}\ \ (\mathrm{IQR}) = 4\ (4)$ дни

Правило на Тюки

Отклоняващи се стойности (аутлайъри) са
$> Q_3 + 1{,}5 \cdot \mathrm{IQR} = 12$ или
$< Q_1 - 1{,}5 \cdot \mathrm{IQR} = -4$ дни.
Тук: **17, 24, 25 дни**.

Как се намират Q_1 и Q_3

Позиция в подредения ред

$$P = 1 + (n-1) \times p$$

$p = 0{,}25$ за Q_1; $p = 0{,}5$ за Me; $p = 0{,}75$ за Q_3

Дробна позиция → интерполация

Ако $P = k + d$ (k — цяла част, d — дробна):

$$Q = X_{(k)} + d \times \left(X_{(k+1)} - X_{(k)}\right)$$

Методът е от учебника (гл. 4). Различните програми използват малко различни правила — на изпита използвайте този.

Пример: пулс на 10 студенти

Подредени: 72, 73, **74, 78**, 80, 82, **84, 86**, 89, 90

- Q_1: $P = 1 + 9 \times 0{,}25 = 3{,}25 \longrightarrow 74 + 0{,}25 \times (78 - 74) =$ **75**
- Q_3: $P = 1 + 9 \times 0{,}75 = 7{,}75 \longrightarrow 84 + 0{,}75 \times (86 - 84) =$ **85,5**
- $\mathrm{IQR} = 85{,}5 - 75 =$ **10,5 уд./мин**

Дни хоспитализация ($n = 22$): Q_1 — позиция 6,25 (6-а и 7-а стойност са 2) → 2; Q_3 — позиция 16,75 (16-а и 17-а са 6) → 6.

Кой показател кога

Признак / разпределение	**Централна тенденция**	**Разсейване**
количествен, приблизително **симетрично**, без силно отклоняващи се стойности	средна аритметична $\overline{X}$	стандартно отклонение S
количествен, **изтеглено** или с аутлайъри	медиана Me	интерквартилен размах IQR
качествен, номинален	честоти и относителни дялове; мода	—
качествен, ординален	медиана, мода; дялове по категории	квартили

Преди да изберете показател — погледнете разпределението (точкова диаграма, хистограма, боксплот).

СЪДЪРЖАНИЕ

1. Признаци и скали за измерване
2. Показатели за централна тенденция
3. Показатели за разсейване
4. **От извадката към популацията: доверителен интервал**
5. Задача в час и домашна работа

Статистика, параметър и стандартна грешка

Извадка (статистика)	Популация (параметър)
$\overline{X}$	μ
$\hat{p}$	P
S	σ

Параметърът е неизвестен; статистиката го **оценява**. Всяка нова случайна извадка дава малко различна оценка — **случайна извадкова грешка**.

Стандартна (средна) грешка

$$S_{\overline{X}} = \frac{S}{\sqrt{n}}$$

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

(за $\hat{p}$ в проценти; в листа с формули — с дроби: $\sqrt{p(1-p)/n}$)

По-голямо S $\longrightarrow$ по-голяма грешка. По-голямо n $\longrightarrow$ по-малка; за **2 пъти** по-малка грешка — **4 пъти** по-голяма извадка.

Стандартно отклонение ≠ стандартна грешка

Стандартно отклонение S

Колко се различават **отделните пациенти** един от друг. Не намалява с увеличаване на n.
Преживяемост: $S = 3{,}70$ месеца.

Стандартна грешка $S_{\overline{X}}$

Колко точно сме оценили **средната**. Намалява с увеличаване на n.
Преживяемост: $S_{\overline{X}} = 3{,}70/\sqrt{30} = 0{,}675$ месеца.

При $n > 1$ стандартната грешка винаги е по-малка от стандартното отклонение.

Доверителен интервал: формули

оценка $\pm$ критична стойност $\times$ стандартна грешка

Средна аритметична μ

- σ неизвестно (почти винаги):

$$CI = \overline{X} \pm t_{\alpha/2,\,\mathrm{df}} \cdot S_{\overline{X}}, \quad \mathrm{df} = n - 1$$

- σ известно: $CI = \overline{X} \pm Z_{\alpha/2} \cdot \sigma/\sqrt{n}$; при много голяма извадка $t \approx Z$

Относителен дял P

$$CI = \hat{p} \pm Z_{\alpha/2} \cdot S_p$$

Ниво	Z	t, **напр.** $\mathrm{df} = 29$
90 %	1,645	1,699
95 %	**1,960**	**2,045**
99 %	2,576	2,756

t-таблицата в листа с формули: колоната $t_{0,025}$ е за 95 % интервал. Когато стойността на df липсва в таблицата, използвайте най-близкия по-малък ред.

Предпоставки

Независими наблюдения от случайна извадка; при малка извадка — приблизително нормално разпределение без силно отклоняващи се стойности.

Решен пример: 95 % CI за средната преживяемост

1. $\overline{X} = 250/30 = 8{,}333$ месеца
2. $S = \sqrt{396{,}67/29} = 3{,}698$ месеца
3. $S_{\overline{X}} = 3{,}698/\sqrt{30} = 0{,}675$ месеца
4. σ неизвестно $\longrightarrow t_{0{,}025;29} = 2{,}045$
5. $8{,}333 \pm 2{,}045 \times 0{,}675 \approx 8{,}333 \pm 1{,}381$

Междинните стойности са показани съкратено; изчисленията са с пълна точност.

95 % CI за μ: [6,95; 9,71] месеца

Заключение

Оценената средна преживяемост е 8,33 месеца; 95 % доверителен интервал — от **6,95 до 9,71 месеца**.

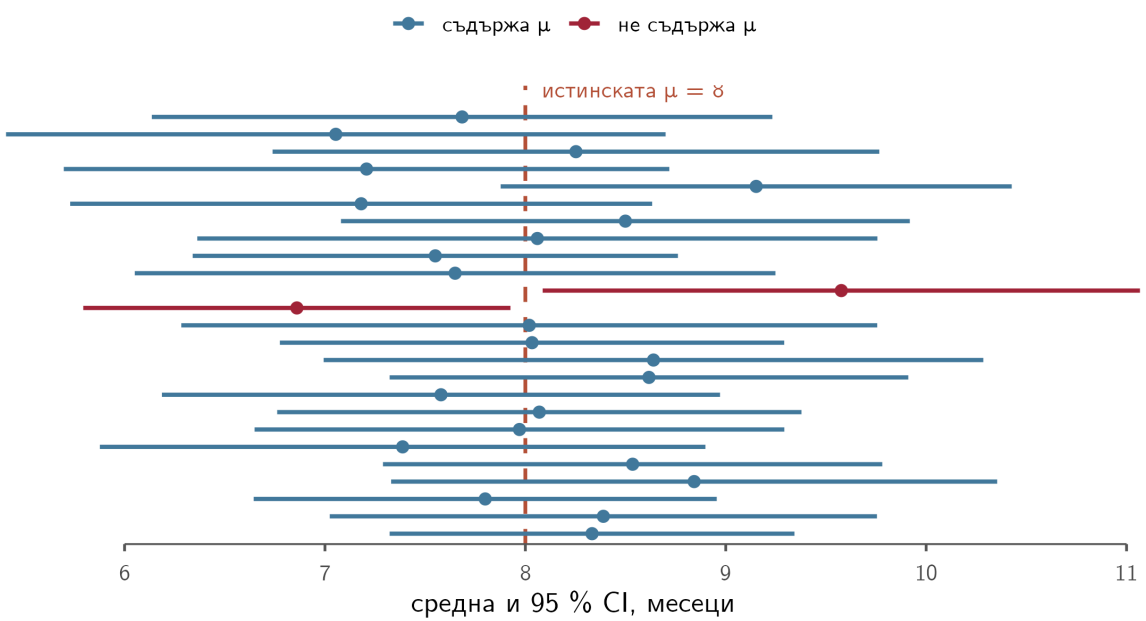
Не закръглявайте рано

С $\overline{X} \approx 8$ и $S_{\overline{X}} \approx 0{,}7$ се получава [6,6; 9,4] — центърът и границите се изместват. От суровите данни: $8{,}50$; [7,19; 9,81].

Какво означава „95 % доверителен“

25 извадки по 30 пациенти от една и съща популация

Всяка извадка дава различен интервал; методът „улучва“ μ в около 95 % от случаите



Правилно

При многократно повторение около **95 %** от построените така интервали съдържат истинската μ.

Неправилно

- „95 % от пациентите живеят 6,95–9,71 месеца.“
- „μ се мени между 6,95 и 9,71.“

Тук 2 от 25 интервала не съдържат μ.

Решен пример: 95 % CI за относителен дял

Диета	С подобрение	Без	Общо
безглутенова	56	44	100
контролна	45	55	100

1. $\hat{p} = 56/100 \times 100 = 56\%$
2. $S_p = \sqrt{56 \times 44/100} = \sqrt{24{,}64} = 4{,}964\%$
3. $56 \pm 1{,}96 \times 4{,}964 = 56 \pm 9{,}729$

95 % CI за P: [46,27; 65,73] %

Кога е приложимо

Нормалното приближение е добро, ако и броят **със**, и броят **без** събитието са поне 5 (учебно правило; по-строго — поне 10). Тук 56 и 44. При малки извадки и редки събития — интервал на Уилсън (със софтуер).

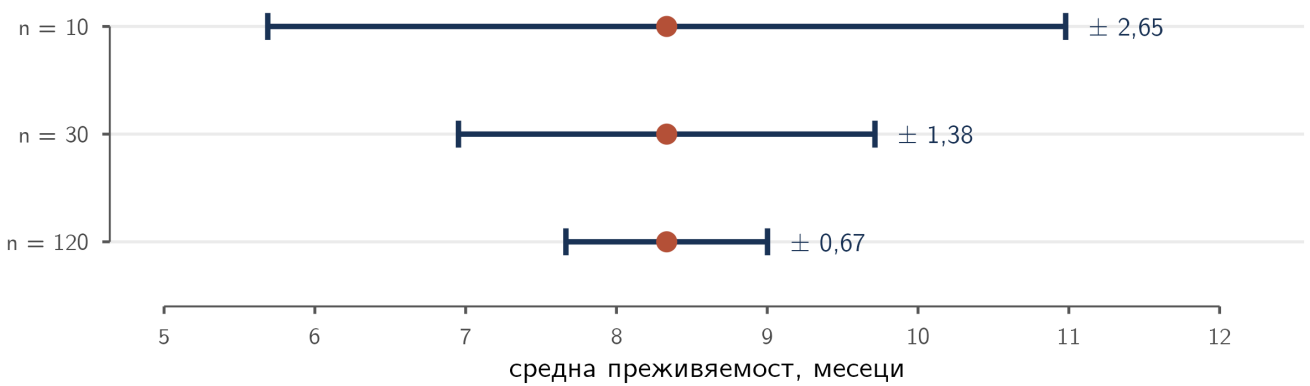
99 % CI

$56 \pm 2{,}576 \times 4{,}964 =$ **[43,21; 68,79] %** — по-широк.

Какво променя ширината на интервала

По-голяма извадка → по-тесен интервал

Същите средна 8,33 и S = 3,70; променя се само n (и t)



Ако...	**ширината...**
$n \uparrow$	**намалява**
$S \uparrow$	**нараства**
99 % вместо 95 %	**нараства**
90 % вместо 95 %	**намалява**

Чести грешки

Грешка	**Правилно**
S вместо $S_{\overline{X}}$ във формулата за CI	интервалът се строи със **стандартната грешка**
$Z = 1{,}96$ при малка извадка и неизвестно σ	**t** с $\mathrm{df} = n - 1$
ранно закръгляне (8,33 $\longrightarrow$ 8)	закръглявайте само **крайния** резултат
смесване на 0,56 и 56 % в една формула	**или** проценти, **или** дроби — навсякъде
„95 % от пациентите са в интервала“	CI е за **параметъра**, не за отделните хора
препокриващи се CI $\longrightarrow$ „няма разлика“	нужен е тест или CI на разликата

СЪДЪРЖАНИЕ

1. Признаци и скали за измерване
2. Показатели за централна тенденция
3. Показатели за разсейване
4. От извадката към популацията: доверителен интервал
5. **Задача в час и домашна работа**

Задача в час: тегло на спортисти

След тренировъчен период е измерено теглото на 8 спортисти (приемете случайна извадка от приблизително нормално разпределена популация):

64, 76, 68, 84, 80, 72, 74, 72 кг

Изчислете средното тегло, медианата, стандартното отклонение, стандартната грешка и **95 % доверителен интервал** за средното тегло в популацията. Формулирайте заключение.

Подсказки

- $n = 8$, $\mathrm{df} = 7 \longrightarrow t_{0,025;7} = 2{,}365$
- направете таблица с $X_i - \overline{X}$ и $\left(X_i - \overline{X}\right)^2$
- проверка: $\sum\left(X_i - \overline{X}\right) = 0$

Решение

- $\overline{X} = 590/8 = \mathbf{73{,}75\ кг}$; подредено: 64, 68, 72, **72, 74**, 76, 80, 84 $\longrightarrow \mathrm{Me} = \mathbf{73\ кг}$
- $\sum \left(X_i - \overline{X}\right)^2 = 283{,}5$
- $S = \sqrt{283{,}5/7} = \sqrt{40{,}5} = \mathbf{6{,}36\ кг}$
- $S_{\overline{X}} = 6{,}364/\sqrt{8} = \mathbf{2{,}25\ кг}$
- $73{,}75 \pm 2{,}365 \times 2{,}25 = 73{,}75 \pm 5{,}32$

95 % CI за средното тегло: [68,43; 79,07] кг

Малка извадка $\longrightarrow$ голямо t (2,365 срещу 1,96) и широк интервал. Методът предполага приблизително нормално разпределение на теглото.

Обобщение и домашна работа

Да запомним

1. Скалата определя допустимите показатели.
2. Симетрично $\longrightarrow \overline{X} \pm S$; изтеглено $\longrightarrow \mathrm{Me}$ (IQR).
3. S — разсейване на хората; $S_{\overline{X}} = S/\sqrt{n}$ — точност на средната.
4. CI = оценка $\pm$ критична стойност $\times$ стандартна грешка; за средна — **t** с $n-1$ степени на свобода.
5. 95 % е надеждността на **метода**.
6. Закръглявайте само накрая; пишете мерната единица.

Домашна работа

Две задачи — на сайта, „Медицинска статистика“: **kostadinoff.github.io**

- Задача 1: дни до възстановяване (интервален ред, $\overline{X}$, S, 95 % CI; $\pm 2S$ срещу CI)
- Задача 2: ваксинационен обхват (95 % и 99 % CI за дял)

Следващия път: **проверка на хипотези, t-тест**. В края — текуща проверка (1 т.) по **доверителен интервал**.

Текуща проверка (1 точка): пряка стандартизация

Правила

- **15 минути**, самостоятелно
- един казус на човек; съседите имат **различни** варианти
- позволени: калкулатор и лист с формули
- напишете име, факултетен номер и група

Какво се оценява

1. специфични показатели по групи
2. стандарт (структура в дроби, сбор 1)
3. стандартизирани показатели
4. **заключение с изречение**: коя група е с по-висок показател и защо изводът се променя

Благодаря за вниманието!