

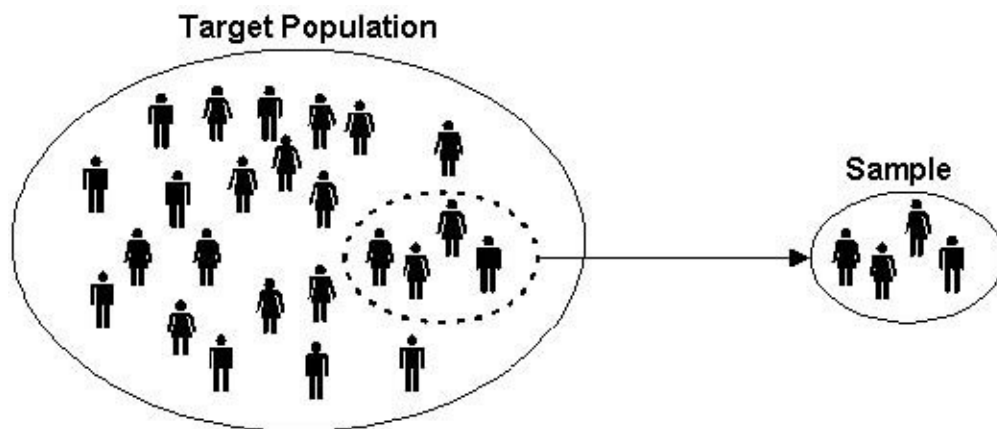
02–Упражнение 2025/2026

Дескриптивна и инферентна статистика

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

Увод

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



Фиг. 1: Генерална съвкупност и извадка

¹ За да направим това „обобщение“ е необходимо извадката да бъде **представителна** – тоест, освен да бъде случайно подбрана, извадката трябва да възпроизвежда структурата (най-малко по пол и възраст) на генералната съвкупност.

Основни понятия

Важно!

*Дескриптивната (описателна) статистика представлява анализ (описание) на статистическите явления **в извадка**. Инферентна статистика е процесът на генерализация (обобщение) на получените резултати в **генералната съвкупност***

Статистически променливи

Статистическите данни представляват числови или буквени стойности, записани след измерване или изброяване на определена характеристика върху група от обекти (пациенти). Тъй като измерванията могат да приемат различни числови или нечислови стойности, те се наричат още **променливи**. За една група студенти от втори курс например, може да се установи, че се различават по пол, възраст, успех, отношение към определен проблем и редица други **характеристики**. Всички посочени могат да бъдат наречени **променливи**. От друга страна, ако характеристиката е една и съща за всеки член на групата, тя се нарича **константа**.

Статистическият признак

Определение!

Статистическият признак представлява отличителна черта, белег на една или повече единици – обекти на статистически изследвания.

Статистическите признаци се различават по вид, което налага различен подход при статистическото им изучаване. Основното им разделяне е на **вариационни** (още количествени) и **категорийни** (още качествени).

Вариационни са тези признаци, които се изразяват с **реални числа**. Те могат да бъдат **прекъснати (дискретни)** и **непрекъснати (континуитетни)**. Прекъснатите могат да приемат отделни, изолирани една от друга числови стойности. Такива са: броят на членовете на отделно домакинство, броят боледувания за една година, броят на живородените деца и др.² Непрекъснати са признаците, които могат да бъдат измерени с голяма точност след десетичната запетая. Такива са: теглото, температурата, кръвното налягане.

² Въпреки, че прекъснатите характеристики не се изразяват с дробни числа, много често това се налага, когато те се анализират като прекъснати променливи. Така, въпреки че за някого звучи абсурдно, може да кажем, че средностатистическата българка ражда 1.7 деца.

Категорийните признаци са тези, чиито характеристики нямат самостоятелен числов израз. За тях се използват „етикети“ или описателно буквено наименование. Такива са например пол, семейно положение, кръвна група, вид заболяване и др.

Скали за измерване

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

Таблица 1: Видове признаци и скали на измерване

Скала	Характеристика	Пример
Номинална	Категории, „наименования, етикети“.	Диагноза, кръвна група
Дихотомна	<i>Вид номинална скала за признаци само две възможни стойности.</i>	<i>пол</i>
Ординална	Качествени променливи, които имат логически (нарастващ или намаляващ) ред. Не е възможно изчисление на пропорция или разлика	Степен на тежест на сърдечна недостатъчност (лека, умерена, тежка), стадии на онкологично заболяване (I-ви, II-ри ..)
Рангова	<i>Вид ординална скала при която единиците се записват с рангове (поредни места) от 1 до n.</i>	<i>Оценки „слаб 2“, „среден 3“ и т.н.</i>

³ Ако кажем за един човек, че е висок, това не е равнозначно да кажем, че неговият ръст е 185 см. В първия случай използваме неточно измерване на „номинална скала“, а във втория – точно на „пропорционална скала“

Скала	Характеристика	Пример
Интервална	Количествени променливи. Стойността „0“ не означава липса на признака, тя е допустима наред с други позитивни и отрицателни стойности. Не позволява отношения	Температура (NB: Ако днес е 10 градуса по-топло от вчера, няма как да изчислим „колко процента по-топло е“)
Пропорционална	Количествени променливи. Стойността „0“ означава липса на признака. Позволява съотношения	Тегло (можем да го измерим с голяма точност след десетичната запетая).

Дескриптивен статистически анализ

Статистическият анализ обхваща математическо проучване посредством подходящи методи на **обобщаващи числови характеристики** и смисловото им интерпретиране. Тези характеристики се разделят на два вида – описващи **централната тенденция** и **разсейването** в изследваната извадка⁴.

Централната тенденция при променливите, които се измерват на **пропорционална** или **интервална** се описва посредством **средната аритметична** и **медианата**. Разсейването при тези променливи се онагледява чрез стойността на **стандартното отклонение** или **интерквartilния размах**.

За променливите измервани на **номинална** и **ординална** скала се използва единствено индикаторът относителен дял.

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

Показатели за централна тенденция

Средна аритметична

Определение!

Средната аритметична се изчислява в зависимост от вида на представяне на количествените данни. Те могат да бъдат изброени *негрупирани числа* или таблично подредени агрегирани и групирани стойности. От своя страна групировката може да бъде извършена във *вариационен* или *интервален* рег.

Негрупирани данни

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

$$\bar{x} = \frac{1}{n} \cdot \sum x$$

Където:

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

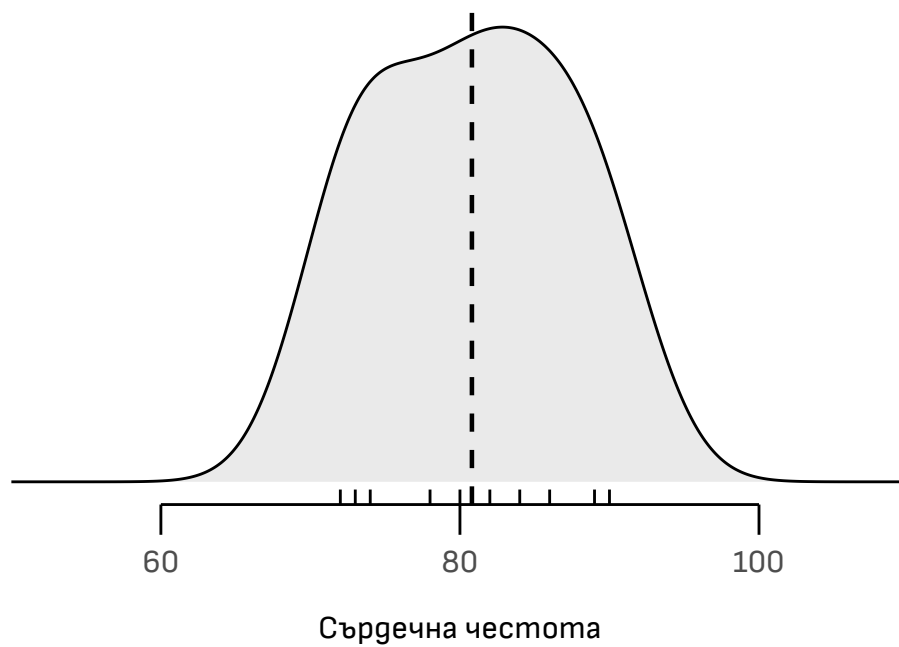
Нека онагледим това с пример. По-долу са представени стойностите на променливата „сърдечен пулс“ за 10 студента от втори курс.

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

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

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

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



Фиг. 2: Графично разпределение на пулса – по хоризонтала са представени наблюдавани стойности, а по вертикала колко често са наблюдавани те в извадката. С черна пунктирана линия е представена средната аритметична. Посоченото разпределение се приближава до нормалното разпределение – има камбановидна, симетрична форма

Групирани данни

Интервален рег

Когато разполагаме с по-голям обем данни, удобно е тяхното представяне в групирана форма.

При **интервалния рег** тази групировка се извършва чрез „разкъсване“ на стойностите на изследваната променлива в равномерни и непривокриващи се стъпки – интервали. Всеки от тях се определя от долна и горна лимитираща стойност (граница). При формирането на интервалните групи се преброяват само участниците, чиято наблюдавана стойност е в границите на интервала. С други думи, при тази групировка „губим“ статистическа точност, защото не знаем индивидуалните стойности на участниците, а само към коя група принадлежат. Пример за интервална групировка е представен в Таблица 2.

Таблица 2: Пример за интервален рег при измерването на пулс сред 45 студенти

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

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

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

Където:

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

• $\sum f$ е сумата от всички наблюдавани участници.

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

Таблица 3: Стъпка 1 – Определяне на средата на всеки интервал

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

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

Таблица 4: Стъпка 2 – Умножение на средата на интервала по броя на наблюдаваните студенти в него

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

Накрая сумираме всички получени произведения и ги разделяме на броя студенти

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

Степенен рег

Вторият вариант за групиране на статистически данни е **степенния рег**. При него записваме всяка стойност на интересуващата ни променлива, а срещу нея нанасяме

броя на наблюдаваните лица, при които сме я наблюдавали. Пример за степенен ред е представен в Таблица 5.

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

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

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

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

Където:

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

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

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

Стойност на пулса	Брой студенти	Произведение $x \times f$
59	1	59

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

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

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

Други показатели за централна тенденция

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

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

⁵ **Важно!** За разлика от интервалния при степенния ред, знаем стойността на всяко едно наблюдение. Тоест не „губим“ статистическа точност.

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

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

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

Относителен дял

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

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

Където:

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

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

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

Показатели за разсейване

Показателите за разсейване са мерки за разпръснатостта на стойностите спрямо централната тенденция. В настоящия курс ще разгледаме показателите за разсейване само за променливи измервани на **пропорционална** или **интервална** скала.

Размах

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

$$R = 90 - 72 = 18$$

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

Стандартно отклонение

Стандартно отклонение при негрупирани данни

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

$$SD = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

Където:

- SD е *стандартното отклонение* – мярка за това колко силно се разпръснати отделните наблюдения около средната стойност. По-голямо SD означава по-голяма вариабилност.
- X_i е стойността на i -тото наблюдение в набора от данни.
- $\bar{X}$ е *средната аритметична* на всички наблюдения
- $(X_i - \bar{X})$ е *отклонението на всяко наблюдение от средната стойност*. Ако получената стойност е положителна → наблюдението има стойност, над средната.
- $(X_i - \bar{X})^2$ е *квадратът на отклонението*. Квадратирането (повдигането на квадрат има две цели: (1.) Премахва знаците (не се „зануляват“ + и – отклонения). По-големите отклонения тежат непропорционално повече → формулата е чувствителна към екстремни стойности.
- $\sum_{i=1}^n (X_i - \bar{X})^2$ е *сумата от квадратите на отклоненията*. Това е общата „квадратична вариация“ в данните.
- $\frac{1}{n-1}$ е *делението на броя на наблюденията минус едно* – т.нар. **Беселов коректор**. То се използва, когато работим с *извадка*, а не с цялата популация, и гарантира, че оценката на вариацията е точна.
- $\sqrt{\cdot}$ е *корен квадратен*, който *връща* стойността обратно в оригиналните единици.

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

Ако съберем отклоненията на всички участници, ще получим нула. Това няма да ни позволи да изчислим средното отклонение в групата – затова, във втората стъпка повдигаме всяка една от девиациите на втора степен. Изчисленията са представени в Таблица 8

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

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

Таблица 8: Девияция повдигната на квадрат

сърд.честота	отклонение	отклонение 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

В третата стъпка – събираме стойностите на всички отклонения повдигнати на квадрат. Резултатът разделяме на броя на студентите, а за да „стандартизираме“ използваното в предходния етап степенуване, използваме реципрочната математическа функция – коренуване.

$$SD = \sqrt{\frac{1}{10-1} \cdot (84.64 + 0.64 + 67.24 + 77.44 + 60.84 + 46.24 + 7.84 + 27.04 + 10.24 + 1.44)}$$

$$SD = \sqrt{\frac{1}{9} \cdot (383.6)}$$

$$SD = \sqrt{42.62} = 6.53$$

След като вече разполагаме със стандартното отклонение, можем да изразим пулса на групата по този начин $\bar{x} = 80.8 \pm 6.53$

Стандартно отклонение при данни групирани в интервален рег

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

$$SD = \sqrt{\frac{\sum (x_u - \bar{x})^2 \cdot f}{n - 1}}$$

Където:

- SD е символът за стандартно отклонение;
- x_u е средата на всеки от интервалите;
- $\bar{x}$ е средната аритметична;
- f е броят на наблюденията за всеки един интервал;
- $x_u - \bar{x}$ е отклонението на средата на интервала от средната аритметична;
- $\sum (x_u - \bar{x})^2 \cdot f$ е сумата от квадратите на всички отклонения, умножени по броя на наблюденията в интервала;
- $n - 1$ е общият брой на участниците, намален с единици (корекция на Базел)

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

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

Таблица 9: Стъпка 2 – Определяне на девиацията

пулс	среда на интервала	Девиация $x_u - \bar{x}$	брой студенти
48 – 52	50	-15.33	4
53 – 57	55	-10.33	5
58 – 62	60	-5.33	8
63 – 67	65	-0.33	11
68 – 72	70	4.67	6
73 – 77	75	9.67	6
78 – 82	80	14.67	5

В трета стъпка – повдигаме девиациите на квадрат, както е показано в Таблица 10.

Таблица 10: Стъпка 3 – Повдигане на девиациите на втора степен

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

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

Таблица 11: Стъпка 4 – Умножаване на квадрата на девиациите по броя студенти

пулс	среда на интервала	$(x_u - \bar{x})^2$	брой студенти (f)	$(x_u - \bar{x})^2 \cdot f$
48 – 52	50	235	4	940
53 – 57	55	106.7	5	533.5

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

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

$$SD = \sqrt{\frac{940 + 533.5 + 227.2 + 1.21 + 130.8 + 562.2 + 1076}{44}}$$

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

Важно

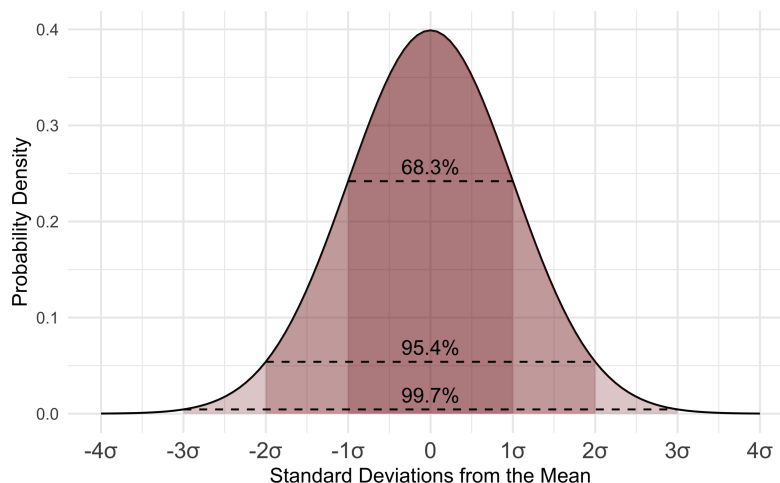
В основата на стандартното отклонение можем да изградим **предикативен интервал**. Това е възможно, само ако изследваната променлива е **нормално** разпределена. С помощта на този предикативен интервал, можем да „предскажем“ какъв брой от изследваните имат стойност в конкретен диапазон от стойности. За целта използваме правилото на **трите сигми**:

- 95 % от наблюденията **в извадката** имат стойност на променливата в интервала между средната аритметична + или – **две** стандартни отклонения;
- 68% от наблюденията **в извадката** имат стойност на променливата в интервала между средната аритметична + или – **едно** стандартно отклонение;
- 99.7% от наблюденията **в извадката** имат стойност на променливата в интервала между средната аритметична + или – **три** стандартни отклонения.

Нормално разпределена величина – означава, че имаме много наблюдения със стойности близки до средната аритметичната, която съвпада с медианата и модата. При построяването на графика – разпределението е камбановидно и симетрично.

*Ако сме изчислили теглото на извадка от 1000 души и сме установили че $\bar{x} = 80 \pm 10$ и приемем променливата за **нормално разпределена** можем да твърдим, че 95 % от участниците (950 изследвани) имат тегло в диапазона $80 \pm 2 \cdot SD = 80 \pm 20$ или от 60 до 120 кг. Останалите 2,5 % от останалите 50 участници (25) ще имат тегло над 110 кг, а 25 души (2,5 %) ще имат тегло под 60 кг.*

Илюстративно нормалното разпределение е представено на фигура Фиг. 3 .



Фиг. 3: Нормално разпределение

Интерквартилен размах

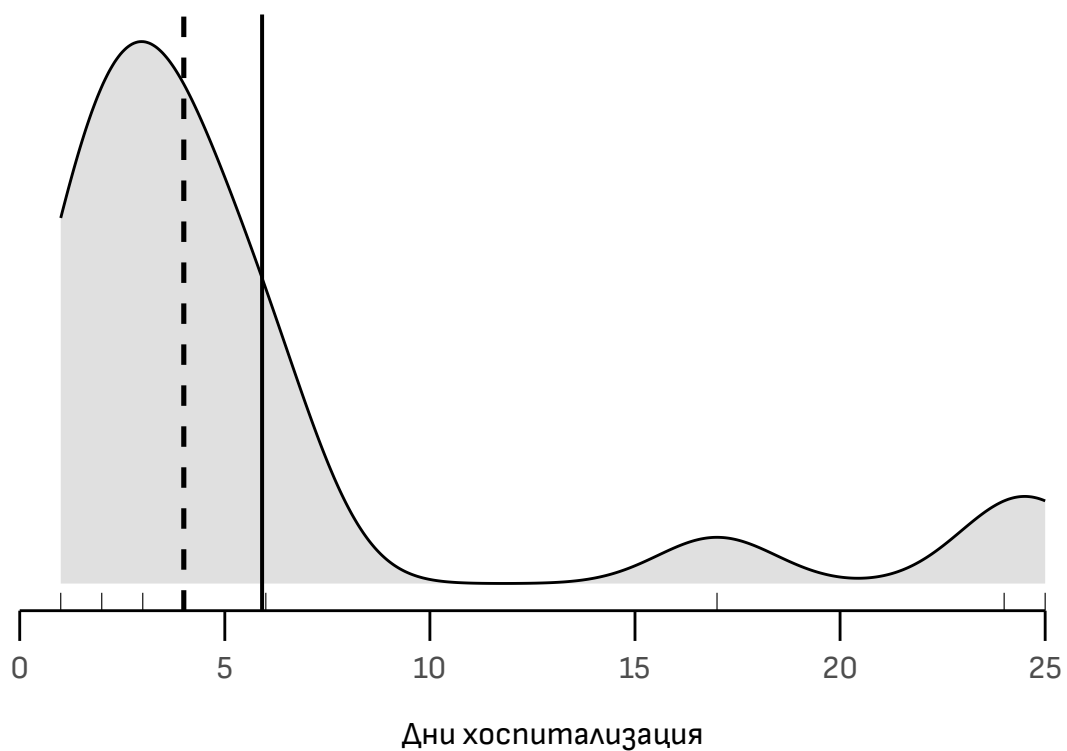
Понякога средната аритметична не е подходяща за измерване на централната тенденция. Такъв случай е разгледан по-долу, когато използваме данните за продължителността на хоспитализацията на 24 пациенти.

[1] 24 4 2 3 4 1 6 17 4 3 2 6 25 1 6 4 2 4 3 6 2 1

По-голямата част от включените в статистическия ред пациенти са прекарвали между 1 и 6 дни в болница. Въпреки това, ако изчислим средната аритметична за цялата група получаваме 5.91 дни. Това число не представя добре централната тенденция, тъй като само 3 от 24-те (12.5%) пациенти са пролежали в болница за повече от 6 дни. Техните стойности се наричат екстремални, а средната аритметична силно повлияна от тях.

Ако нанесем броя на дните по хоризонтала, а броя на пациентите по вертикала ще получим разпределението на променливата „*продължителност на хоспитализацията*“ – тя е представена на Фиг. 4.

Представеното в Фиг. 4 разпределение **ясно изтеглено**. С други думи, по-голямата част от наблюденията се намират в лявата част на средната аритметична, а пациентите с екстремални стойности издърпват „*опашката*“ на разпределението на дясно. В случаи като тези, медианата е по-добър показател за централна тенденция и нейната стойност, по правило е по-малка от средната аритметична. Обратно, при **ляво изтеглено разпределение**, повече стойности са са високи и се намират в дясната половина на разпределението, единични екстремни ниски стойности издърпват „*опашката на разпределението*“ наляво. В тези случаи стойността на медианата е по-голяма от стойността на средната аритметична. При **нормално** разпределение, медианата и средната имат приблизително или изцяло идентични стойности.



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

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

$$IQR = Q3 - Q1$$

Където:

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

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

[1] 1 1 1 2 2 2 2 3 3 3 4 4 4 4 4 6 6 6 6 17 24 25

Ако разделим тези 24 пациенти в 4 групи (по 6-ма) ще установим стойностите на 4-те квартила. В случая 6-тият пациент, пролежал 2 дни, е на границата между първата и останалите три четвъртини на статистическия ред. Стойността, на неговото пролежаване представлява 1-вия квартил $Q1$. Стойността на пролежаването на 12-тия пациент (4 дни) ограничава разделя статистически ред на две равни части, това е и медианата или втория квартил $Q2$ – половината пациенти са пролежали повече дни от него. Стойността на пролежаването на 18-тия пациент (6 дни) е 4-тия квартил $Q3$.

След като сме определили квартилите прилагаме формулата. Интерквартилният размах е

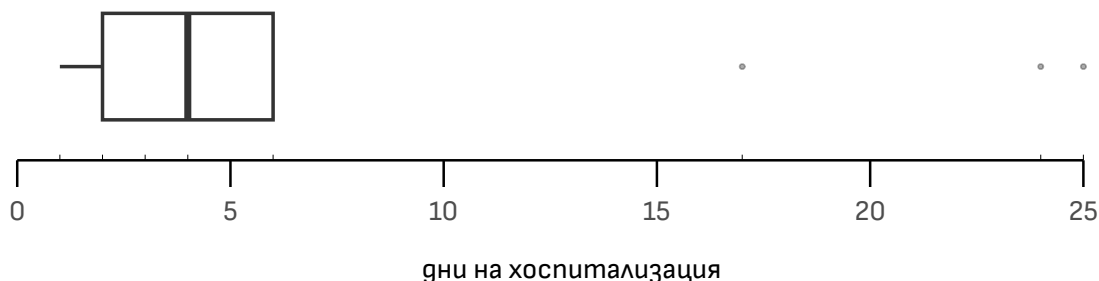
$$IQR = Q3 - Q1 = 6 - 2 = 4$$

За графично представяне на интерквартилният размах използваме боксплот диаграмата представена на Фиг. 5.

С помощта на интерквартилният размах можем да определим кои стойности са „екстремални“ ¹¹

¹⁰ „Квартилите“ са стойностите на наблюденията, която разделят статистическия ред на 4 равни части.

¹¹ Екстремални са всички стойности по-високи от $Q3 + 1.5 \cdot IQR$ или по-ниски от $Q1 - 1.5 \cdot IQR$. В посочения пример такива са стойностите над $6 + 1.5 \cdot 4$ и под $2 - 1.5 \cdot 4$.



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

Инферентна статистика

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

Статистиката изчислена в извадка се използва за да се установят **вероятните граници** на параметъра в генералната съвкупност

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

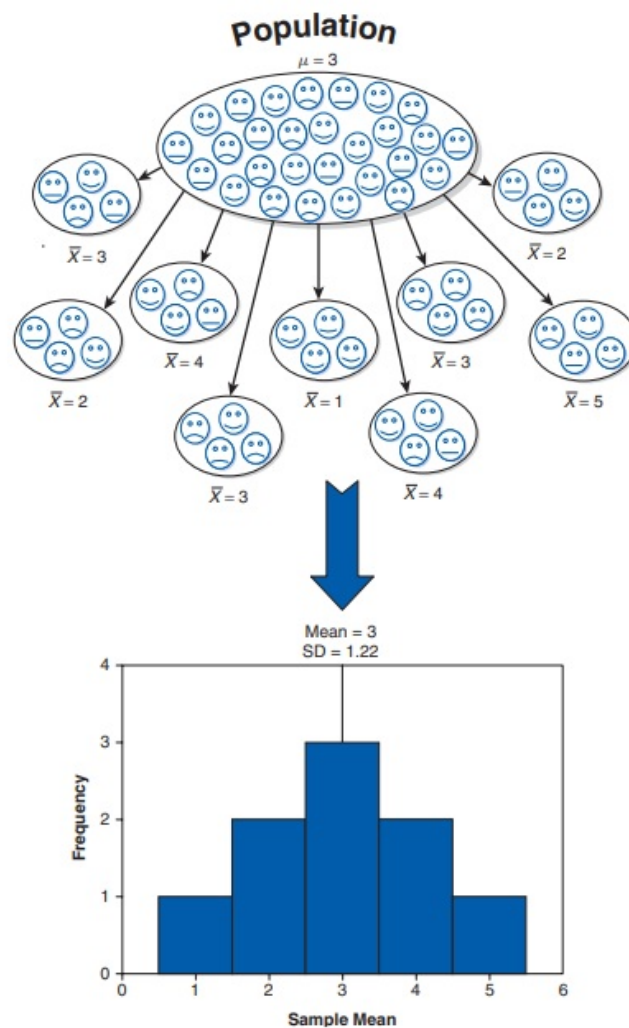
Стандартна грешка на средната аритметична (SEM)

Нека си представим, че имаме генерална съвкупност от 200 студенти със стойност на средната аритметична равна на 3. Ако направим 9 случайни извадки (всяка по 10 студенти), можем да определим средната аритметична във всяка една от тях. Тази извадкова средна ще се различава с малко или повече от истинската поради случайната грешка, която не можем да избегнем. Ако нанесем тези 9 извадкови средни статистики на една графика ще получим **нормално разпределение** (Фиг. 6).

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

За да изчислим стандартната грешка прилагаме формулата:

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



Фиг. 6: Разпределение на средни аритметични на 9 извадки от генералната съвкупност

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

Където:

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

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

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

$$SE_m = \frac{SD}{\sqrt{n}} = \frac{6.53}{\sqrt{10}} = \frac{6.53}{3.16} = 2$$

Във втория пример, с данните за 45-мата студенти в интервален ред със средната аритметична 65.33 уд./мин и стандартното отклонение 8.88 уд./мин. Стандартната грешка е

$$SE_m = \frac{SD}{\sqrt{n}} = \frac{8.88}{\sqrt{45}} = \frac{8.88}{6.71} = 1.31$$

Интервал на доверителност средна аритметична

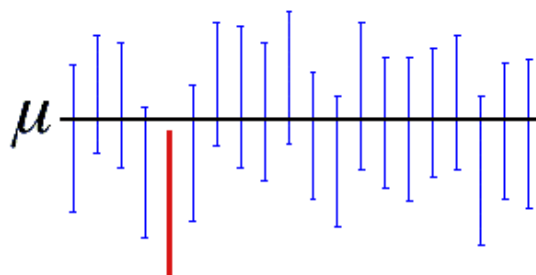
Интервалът на доверителност представлява *оценка* на параметъра на генералната съвкупност. Той ограничава възможните стойности на неизвестния параметър в определен интервал от числа с определена вероятност (сигурност). Тази степен на сигурност се определя от изследователя – в медицината най-често се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

За да се онагледят концепцията за интервал на доверителност се използва Фиг. 7.

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

За да конструираме интервала на доверителност използваме формулата:

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



A 95% confidence interval indicates that 19 out of 20 samples (95%) from the same population will produce confidence intervals that contain the population parameter.

Фиг. 7: 95% интервал на доверителност сред 20 извадки

Където

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

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

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

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

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

Стандартна грешка и интервал на доверителност за пропорция

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

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

⚠ Важно!

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

Първо – изчисляваме относителния дял (статистика) в извадката

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

След това изчисляваме стандартната грешка

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

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

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

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

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

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

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

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