02-Упражнение

Дискриптивна и инферетна статистика

ас.д-р Костадин Костадинов

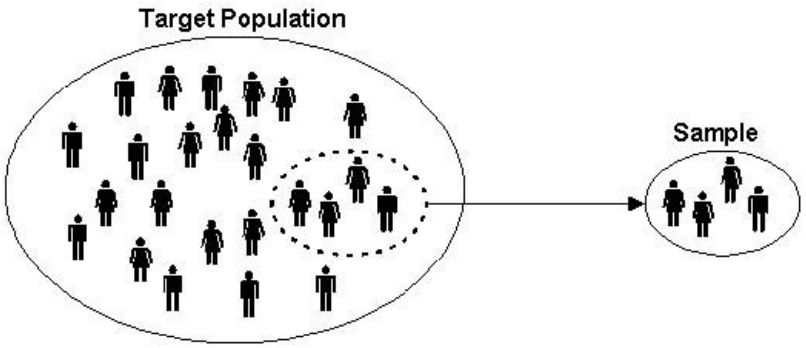
Вместо увод

Представете си 30 милиона лева накуп. Толкова е бюджетът на най-мащабното статистическо изследвне организирано миналата година от Националния статистически институт. Това е преброяването на населението и жилищния фонд. Организацията на подобно изследване е огромна задача. Изисква над 27 хиляди преброители, поне 2 IT фирми, една дузина статистици и достатъчно време за да обясним с кого живеем или колко е голяма кухнята ни. *Защо обаче са необходими тези данни?* Всъщност причината е известна още от древността. Данните от преброяването се използват от държавната администрация в процеса на вземане на решения и разпределяне на финансиране в различни области като образование, здравеопазване, социална защита и други. Данните от преброяването са в основата на разработването на национални и регионални политики. # Основни понятия

Изследване от подобен мащаб, често е невъзможна задача пред учените. Представете си, че искаме да изследваме сърдечната функция при пациенти със захарен диабет - за целта няма да е необходимо да изследваме само български пациенти, ще трябва да изследваме всички диабетици на планетата - това е практически невъзможно. Необходими са огромни ресурси с които неразполагаме. Тези ресурси са **финансови ресурси** - ще са необходими много повече от 30 милиона лева, **човешки ресурси** - тук няма да имаме нужда от 27 хиляди “преброители”, а много по-голям екип и то лекари, които да извършат изследването, а не на последно място

времеви ресурс - ако за преброяването е необходимо около година подготовка и месец работа “на терен”, то за подобно изследване ще са ни необходими години подготовка и още няколко по толкова за самото изследване.

Как тогава можем да знаем изобщо нещо в медицината? Никой не изследва всички хора, никой не тества лекарства върху всички пациенти (това освен невъзможно, е неетично и незаконно). Отговорът разбира се, чрез статистиката. Статистика ни позволява да изследваме сравнително малка част от хората, част от групата представляваща интерес за изследването. Хората попаднали в изследваната група се наричат извадка, а интересуващата ни група - генерална съвкупност. В основа на резултатите, които наблюдаваме в извадката, можем да направим извод за генералната съвкупност. Това е възможно, само ако хората в тази извадка са избрани случайно (иначе казано сме използвали рандомизация).



В някои изследвания, не се хора, а експериментални животни, затова използваме понятието единици на наблюдение за да отличим участниците в една извадка.

Определения

> ⚠ Дискриптивна и инферентна статистика
>
> Дискриптивна (описателна) статистика е процеса на анализ на явленията вътре в изследваната извадка. Инферентна статистика разбираме процеса, при които правим извод (обобщение) за генералната съвкупност,

в основа на изследваните показатели в случайната извадка.

Видове признаци

Нека си представим едно хипотетично изследване. Една изследователка хипотеза може да твърди, че студентите са под огромен стрес по време на упражненията по статистика. *Но как можем да измерим стреса?* Като показател може да използваме сърдечната честота на всеки студент по време на упражнения. В случая сърдечната честота е **признак на наблюдение.**

⚠ Статистическият признак на наблюдение

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

Може да мислим за признаците малко по-абстрактно. Всичко около нас (а и в нас) може да се опише с определени характеристики. За да опишем един студент можем да използваме неговата височина, тегло, успех от от положените изпити до момента, пол, сърдечната му честота и т.н. Всички посочени характеристики са признаци на наблюдение.

⚠ Признаците са два основни типа: количествени и качествени.

Количествени са тези характеристики на изследвания обект, които могат да бъдат изразени числено. Например, ръстът на студента може да бъде - 173 см (числено изразяване на признакът ръст). **Качествени** са тези характертистики на обекта, които нямат самостоятелно числено значение. Например полът на студента, обект на нашето изследване, може да приема стойностите “мъжки” или “женски”. Кръвната група също е пример за качествен признак, който няма “числено” значение.

Скали за измерване

Качествени признаци

Освен че признаците се разделят на количествени и качествени, те трябва да бъдат измерени. За измерването на тази признаци се използват различни скали. Тези скали определят "типа" на данните, които събираме. Това е от особено значение за последващия статистически анализ върху тях!

Такава скала е **номиналната**. Признаци като "диагноза", "кръвна група", "зодиакален знак" и прочие се измерват на номинална скала. Тази скала измерва качествени признаци, които имат отделно значение. Измерването на номинална скала представлява "етикет" на събраните данни.

Някои признаци могат да вземат само две стойности- полът със стойности "мъжки" или "женски", или наличието на заболяване със стойности "болен и здрав". В тези случаи, измерваме тези качествени признаци се осъществява на т.н **"дихотомна скала"**. Някои статистици, наричат тези признаци с две възможни стойности **алтернативни**.

Ако съберем данни за кръвната група на всички студенти в една група това представлява измерване на номинална скала. За всеки студент, поставяме "етикет" обозначаващ качествената характеристика кръвна група

В медицината, често използваме термини, като *леко*, *умерено* и *тежко състояние*, когато предоставяме информация за някои пациент[1]. Такива признаци на наблюдение , чиито стойности, могат да се съизмерват помежду си, но нямат конкретно числово изражение, се измерват на **ординална скала**.

[1] В този случай признакът на наблюдение е състоянието на пациента. Признакът е качествен, защото не можем да кажем, че *умерено тежко* състояние е 2 или 3 пъти "по-добре" от *тежко* състояние. Въпреки това тези стойности, етикети, с които описване състоянието на пациента имат определена логическа подреба.

Количествени признаци

Те могат да бъдат измервани на интервална и пропорционална скала.

Интервалната скала, съдържа както положителни, така и отрицателни числа и стойността "0" не означава отсъствие на признакът на наблюдение, тази стойност е възможна и се намира в рамките на логиката на измерване на този признак.[2].

[2] При измерване на температурата по скалата на Целзий, нулева е температурата на топене на леда. Но 0 градуса, не означава липса на температура въобще.

При пропорционалната скала “0”-лата има значение на “старт”. Това означава, че стойност “0” е пълно отсъствие на измервания признак[3]. При пропорционалната скала, отношението на всеки две стойности от скалата, не зависи от единицата на измерване. Това свойство дава възможност не само да се сравняват разлики между обектите, но и да се разглеждат отношения между тях [4]. Тоест тази скала, ни позволява да изчисляваме пропорции, да сравняваме не само разликата преди и след експеримент, но и да сравним и съотношението им.

[3] Ако измерваме сърдечната честота на пациента и отчетем “0”, това на практика означава, че пациента няма сърдечна честота.

[4] Нека приемем,че измерваме сърдечната честота на студентите преди и по време на упражнение по статистика. Честотата на студента Иван е 60 уд/мин преди упражнение, а по време стига до 90 уд/мин, докато сърдечната честота на Мария е 80 уд/мин, преди упражнение и достига до 120 уд/мин по време на занятието. В случая сърдечната честота е количествен признак който измерваме на пропроционална скала. Това ни позволява да кажем, че и двамата студенти показват честотата си еднакво с 50%.

Сила на скалите: „Един признак, няколко скали”

Скалите могат да бъдат преобразувани една в друга. Например, теглото може да бъде измерено на пропорционална скала: можем да посочим с доста голяма точност след десетичната запетая колко тежи един човек. Така измерено с величаната тегло можем да извършим редица статистически операции: можем да изчислим каква е сумата на теглото за всички участници или средната аритметична стойност. Тази пропорционална скала, може да бъде преобразувана в друга. Вместо да измерваме теглото в килограми, можем да използваме *рангове*, така че вместо да записваме килограми, можем да запишем най-тежкия изследван в извадката с ранг “1”, следващият по тегло ще запишем с ранг “2” и така до най-слабият. Такова записване обаче не ни позволява да изчисляваме сумата на теглото, защото вече нямаме конкретната стойност на количествения признак. Всъщност сме преобразували една пропорционална скала в ординална. Можем да продължим да преобразуваме скалите, и вместо да записваме ранг, да запишем етикет “добре охранен”, “дебел”, “клечъв”, “мускулест” и т.н, тези етикети са записвани на номинална скала , тук вече сме загубили дори възможността и за “естествена подредба”. Не можем да намерим нито сумата на теглото, нито средната аритметична, можем само да кажем, кой “етикет” се среща най-често.

Възможността да се преобразуват скалите определя тяхната сила. Номиналната скала се определя като слаба защото не може да се преобразува с друга. Интервалната и

пропорционалната скала са силни защото могат да се преобразуват в други

Скала	Характеристика	Пример
Номинална (дихотомна)	Категории, ”наименования”. Само две възможни стойности.	Вид диагноза, кръвна група Пол
Ординална	Качествени променливи, които имат логически (нарастващ или намалящ) 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 ..)
Интервална	Качествени променливи. Стойността ”0” не означава липса на признака. Допустими са както отрицателни, така и позитивни 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. Съдържа абсолютната стойност на ”0” - липса на признака на наблюдение.	Тегло (можем да го измерим с голяма точност 10 кг; 10,5 кг, 10,54 кг и т.н. ; стойността ”0” означава липса на тегло). 50 кг е с 10 кг повече от 40 кг., тези 10 кг са с 20% повече от 40 кг.)

Показатели за централна тенденция

Средна аретметична.

Негрупирани данни

Нека отново разгледаме хипотетичночното изследване със сърдечната честота на студентите. Понеже, не позлваме всички студенти в генералната съвкупност, за да можем да направим извод за целия курс, ще използвам само една група. Ще приемем, че в учебен отдел разпределят студентите по групи по изцяло случаен принцип. Така

тази изследвана студентска група е случайна извадка на генералната съвкупност. Нека приемем, че сме записали стойностите на пулса по време на упражнение. В случая това са нашите изходни данни.

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

Както прави впечатление, тези стойности са негрупирани. Имаме само 10 наблюдения изброени едно след друго. В случая пулсът, ще представим, като непрекъснат признак, измерен на пропорционална скала.[5]

За статистиката, особено важно е да създадем модел. Както например, за построяването на нова сграда, първоначално се построява малък макет, които да прилича много на истинската сграда, но в умален мащаб. По същият начин, за да установим пулса на целия курс, трябва да създадем модел, върху данните от една единствена група, като приемем, че тя е случайна извадка. *Как обаче ще създадем модел?* В случая модел означава, да моделираме една стойност с която да презентираме цялата група. Едно, единствено число, което да представи "средно" колко е пулса на един студент от тази група. За целта използваме т.н средна аритметична

[5] Такива признаци като сърдечна честота, брой дни, брой пациенти, са винаги прекъснати цели числа. Въпреки това, често в статистиката се използват като непрекъснати за да се ползват за статистически цели (като средната аритметична) например в България средно една жена ражда 1.7 деца през целия си живот.

> **Средната аритемична**
>
> представлява мярка за централна тенденция. Когато работим с негрупирани данни средната аритметична представлява сумата от стойностите на една група от числа, разделени на броя на групата.

Когато изследваме средната стойност за извадката използваме символа $\bar{x}$, още можете да го срещнете като xbar[6]. В примера тази средна стройност е равна на сумата ($\sum$) от всички наши наблюдения върху броя на наблюденията (n)

[6] Формула за средна аритметична е

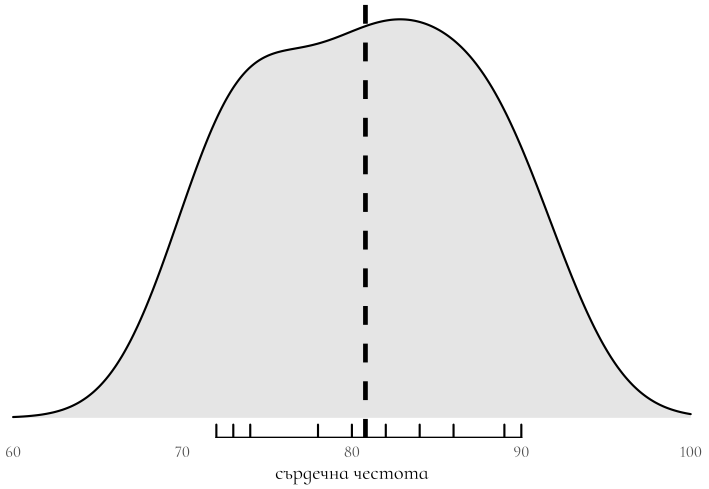
$$\bar{x} = \frac{\sum x}{n}$$

В случая средният пулс е

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

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

Тоест, вече сме създали статистически модел, Ако не знаем точната стойност на пулса на някой студент в групата и някой ни накара да “предположим” някаква стойност, можем да използваме средната аритметична и няма да сме сгрешили много фигура 1



фигура 1: Можем да представим и графично тези данни. На тази фигура - по хоризонтала, ще сложим всички наблюдавани стойности, а по вертикала, колко често сме ги наблюдавали. В резултат ще получим графика, която се нарича емпирично разпределение. С черна пунктирна линия е представена средната аритемична.

Групирани данни

Интервален ред

Понеже сме критикувани за малкия брой единици на наблюдение (едва 10 студенти), няколко колеги желаят да ни помогнат за нашето изследване. Сега вместо 10 студенти от всички 200, имаме данните на 45 случайно избрани студенти от 2-ри курс. Нека разгледаме данните, в таблица 1

Както се вижда, тук данните вече са групирани това на практика означава, че студентите са разпределени в групи спрямо интервала на сърдечната честота. В първата група са студените с сърдечна честота от 48 до 52 уд/мин. Стойности в този диапазон са наблюдавани сред 4-ма студента. В следващият интервал са 5-ма студенти с честотата в диапазона между 53 и 57 уд/мин. Тук е важно да се подчертае някои

таблица 1: Пулс, сред 45-студента.

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

основни елементи на статистическото групиране. Това групиране на данните се нарича още групиране в интервален ред.

- Интервалите имат еднаква ширина. Това е изпълнено в случая: интервалите са през 4 удъра в минута.
- Интервалите не се застъпват. Това също е изпълнено, първият интервал започва от 48 уд/мин до 52 уд/мин, вторият интервал започва с 53 уд/мин. Това означава, че ако имаме студент с 52 уд/мин, ще го преброим само в първият интервал.
- Не знаем точния пулс на участниците. Това е един от недостатъците на групирането в интервални редове: след като погледнем таблицата можем със сигурност да кажем, че имаме 11 участника с пулс между 63 и 67 уд/мин. Въпреки това, ние не знаем точната стойност на пулса за всеки един от участниците. Това **намаля точността**, с която може да определим средната аритметична.
- Интервалният ред е подходящ при голям обем данни, позволява по-бързо пресмятане на интересуващите ни показатели.

Но как изчисляваме средната аритметична в този случай? Отново разполагаме с формула, този път тя изглежда в вида [7]. За да изчислим средната аритметична $\bar{x}$ трябва да умножим x_u по f. x_u? е средата на интервала. В нашият пример първият интервал съдържа стойностите от 48 до 52. Неговата среда ще открием, като съберем 48 и 52 и ги разделим на 2 - получаваме 50 - това е средата на първия интервал:

[7]
$$\bar{x} = \frac{\sum x_u f}{\sum_f}$$

Трябва да извършим този процес за всички интервали, така получаваме таблица 2.

f е наблюдаваната честота. Иначе казано сред колко студента сме наблюдавали в този интервал. След това формулата изисква от нас да умножим x_u по f. За първият ни интервал това означава да умножим средата 50 по броя на студентите в него - 4. Получаваме 200

Това умножение, ще направим за всички интервали и ще добавим нова колона с произведението:

таблица 2: Среда на интервала

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

таблица 3: Интервал на пулса, среда на интервала, честота и произведение

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

След като сме направили и тази стъпка, следва да намерим сумата от всички произведения - $\sum$. За да го направим трябва да съберем всички числа намиращи се в третата колона. В резултат получаваме 2940. Това е стойността, която трябва да поставим в числителя. За знаменател, отново поглеждаме формулата по-горе. Виждаме изписано: $\sum_f$, което означа, че ни е необходимо сумата от наблюдаваните честоти (или броя на всички студенти). Тоест за знаменател ще използваме числото 45 [8].

[8] След тези пресмятания следва да заместим във формулата

$$\bar{x} = \frac{\sum x_u f}{\sum_f} = \frac{2940}{45} = 65.33$$

Степенен ред

Понякога можем да получим отново групирани данни, но не в интервален, а в степенен ред. Да си представим **например**, че сме продължили нашето изследване за пулса, този път сме отчели всяка една стойност на пулса, а срещу нея нанасяне броя на студентите, които сме отчели с нея. Получаваме информацията, както е показано в таблицата.

таблица 4: Степенен ред

Стойност на пулса	Брой студенти
59	1
60	0
61	2
62	4

Стойност на пулса	Брой студенти
63	0
64	3
65	2
66	1
67	1
68	3

Отново имаме формула, която да ни подскаже, как да изчислим средната аритметична, при такъв тип групиране на данните.[9]:

В случаят, трябва да умножим всяка наблюдавана стойност на пулса по честотата с която сме я наблюдавали. Обърнете внимание, че в таблицата сърдечна честота 63 уд/мин не е наблюдавана при нито един студент. Тоест такава, честота всъщност не е наблюдавана въобще. Отново е необходимо първо да извършим умножението, а после да извършим сумирането на всички произведения. Накрая ще разделим полученото число на сумата на всички наблюдавани лица.

[9]
$$\bar{x} = \frac{\sum x f}{\sum_f}$$

Други показатели за централна тенденция

Освен средната аритметична съществуват и други показатели, като представят централната (основаната) тенденция в данните ни. Например:

- Медиана: представлява стойността, която се намира в средата на статистическия ред, т.е. тя е онази стойност, за която половината от измерванията са по-малки от нея, а другата половина са по-големи от нея. [10]
- Мода. Модата M_0 е най-елементарният показател на централната тенденция. Тя се определя като стойността с най-голяма честота в разпределението и се намира непосредствено чрез броене.[11]

[10] **Например** измерили сме кръвна на 5 пациенти, стойностите са 6.0, 6.5, **7.0**, 7.5, 8.0. Стойността 7.0 е медиана - тя разделя този ред на две симетрични части. Ако сме измерили кръвната захар на още един пациент и стойностостта е 8,5, вече имаме четен брой стойности: медианата получаваме, като съберем и разделим на двете стойности в средата на интервала тоест 7.5 и 8.0. Медианата ще е 7.75

[11] Данни за систолното кръвно налягане 120,120,130,135,120,140,140. Най-честата стойност е 120 - това е модата в този статистически ред. Възможно е да имаме и две моди - две стойности, които се срещат еднакво често, тогава наричаме разпределението бимодално.

Когато променливите са номинални или ординални скали, като показател можем да използваме процент. [12]

[12] Ако искаме да разберем какъв е относителния дял на момичетата в групата, ще трябва да разделим броя на момичета, върху броя на студентите в цялата група и да умножим по 100. Формулата има вида $\hat{p} = \frac{m}{n}100$.

Показатели за разсейване.

Размах

Размахът представлява разликата между максималната и минималната стойност в изследваната извадка. В нашия пример размаха е 18 уд/мин. Размахът не се ползва рутинно - понеже е зависим само от две стойности, а понякога те могат да бъдат екстремални поради грешка в измерването.

Стандартно отклонение.

Стандартното отклонение е показател за средното ниво на разсейване вътре в изследваната група. В използвания пример студентите в извадката са със среден пулс 80.8. Въпреки това няма нито един студент с точно тази стойност. Всеки изследван се отклонява с няколко удъра под или над тази средна аритемична. Това отклонение представлява разсейването (девиацията) на стойностите спрямо средната аритметична. Това е направено в таблица таблица 5

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

таблица 5: Отклонение на всеки един от изследваните спрямо средната аритемична за групата

Дали като цяло групата се представява добре с една единствена стойност - тази на средната аритметична или не?. За да разберем това е небходимо да разберем средното ниво на отклонение в извадката. Ако съберем девиациите на всички участници, че получим “0”. Затова можем да повдигнем девиацията на квадрат. По този начин ще премахнем отрицателния знак. Таблично резултатите са представени в таблица 6

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

таблица 6: Деивация повдигната на квадрат

След като имаме позитивни стойности - можем да измерим, какво е “средното” ниво на вариабилност за цялата група. За целта ще съберем стойностите на всички отклонения повдигнати на квадрат и ще ги разделим на броя на участниците в групата. За да получим стандартното отклонение [13] получения резултат се коронува.

За да изчислим стандартното отклонение в примера, трябва да съберем всички стойности на девиацията повдигната на квадрат (това е показателя дисперсия), след което ще ги разделим на обема на извадката и накрая ще короноваме резултата:

[13] Формулата за стандартното отклонение SD.

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

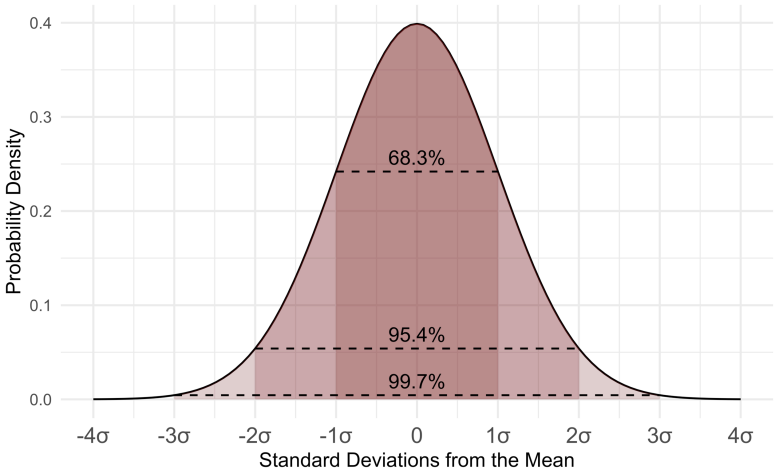
$$SD = \sqrt{\frac{84.64 + 0.64 + 67.24 + 77.44 + 60.84 + 46.24 + 7.84 + 27.04 + 10.24 + 1.44}{10}}$$

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

След като вече разполагаме със стандартното отклонение, можем да изразим пулса на групата по този начин $\bar{x} = 80.8 \pm 6.19$ така записано, вече знаем, средната аритметична за групата, както и стандартното отклонение.

В основата на стандартното отклонение можем да изградим така наречения предикативен интервал. Ако изследваната променлива има **нормално** разпределение, 95 % от студентите в извадката ще имат стойност на пулса някъде в интервала между средната аритметична + или - **две** стандартни отклонения! Нормално разпределена величина - означава, че имаме най-много наблюдения в средата на стойностите и по-малко с отдалечаването от тази среда. Това разпределение предполага и множество статистически пресмятания [14].

Ето и нагледно как да си представите нормално разпределение:



[14] Да си представим, че сме изчислили теглото на извадка от 1000 души. Получили сме следния резултат: $\bar{x} = 80 \pm 10$. Ако теглото е **нормално разпределена величина** можем да твърдим, че 95 % от участниците (тоест 950 от тях) имат тегло в диапазона $80 \pm 2 \cdot SD = 80 \pm 20$ или 60 до 120 кг. Какво остава за останалите 50 - човека. Те представляват 5 % от изследваните. Ако отново следваме закона за нормалното разпределение, това означава, че 2,5 % от тях (25 участници) ще имат тегло над 110 кг и 25 души (2,5 %) ще имат тегло под 60 кг.

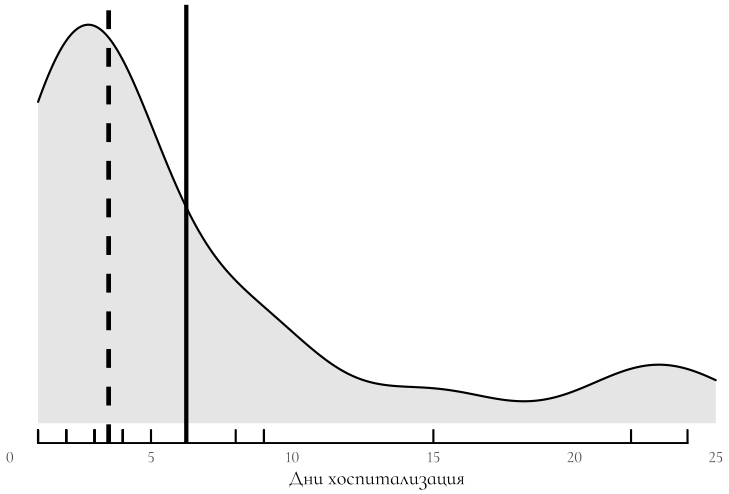
Интерквартилен размах

Понякога обаче, средната аритметична е крайно неудобна за да презентираме цялата група. Нека разгледаме тази данни, които показват колко дни са пролежали пациентите в една болница.

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

Както виждаме, по-голямата част от пациентите са прекарали между 1 и 6 дни в болницата. Ако изчислим средната аритметична получаваме 6.25 дни. Това число обаче не презентира добре хората в извадката. Само 5 от 20 те пациента - вероятно тежко болни, са били хоспитализирани за повече от 6 дни. В случая, средната аритметична е много повлияна от тази пациенти.

Ако нанесем броя на дните по хоризонтала, а броя на пациентите по вертикала ще получим разпределнието на статистическия признак продължителност на хоспитализацията фигура 2. Това разпределение е дясно изтеглено - тоест, по-голямата част от наблюденията (пациенти) се намират в лявата част спрямо средната аритметична, докато само 4-ма "издърпват" опашката на разпределението на дясно.



фигура 2: В този случаи, средната аритметична (непрекъснатата линия) е изместена към тези стойности бегачи (или outliers). По-добър показател на централната тенденция в случая е медианата (средната стойност която разделя всички наблюдения на две части). В случая медианата е 3,5 - прекъснатата линия. Тя по-добре представя средната тенденция в данните.

Когато разпределениетон е изтеглено е по-удобно е да ползваме медианата. В случая медианата е 3 дни. За да разберем какво е отклонението спрямо медианата изпозлваме показателя **интерквартилен размах** Подобно на стандартното отклонение представя какво е разсейването спрямо средната аритметична, така и интеркваритилния размах ползваме, за да оценим каква е вариабилността спрямо медианата.

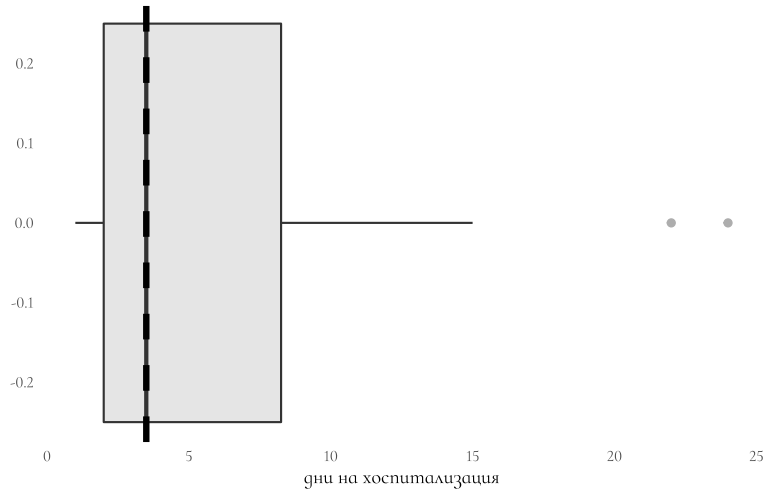
За да разберем интерквартилния размах трбва да подредим пациентите по нарастващ ред на престоя им в болница, така че да започват с най-малката стойност.

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

Ако разделим тези 20 стойности в 4 групи (по 5 пациента) ще получим т.н квартили. В случая 5-тия пациент е пролежал 2 дни, той се намира на границата между първата и оставащите 3/4 от статистическия ред. Това означава, че неговата стойност ще използваме за 1-вия квартил $Q1$ [15]

В средата на реда се намира 10-тия пациент, той разделя реда на две равни части, това всъщност е медианата, той е пролежал 3 дни, затова 2-рия квартил (50% перцентил) е равен на 3. За да намерим стойността на 3-тия квартил (75-тия перцентил) използваме същата логика - трябва да намерил стойността, която има 15-тия участник: това е 4 дни. Преди него се намират стойностите на 75% от наблюденията ни. Тоест неговата стойност е 4-тия квартил това е 4.

Сега можем да намерим и интерквартилния размах IQR. Това е разликата между първия квартил (25 тия перцентил) и третия квартил (75 тия перцентил). Тоест изваждаме 2 от 4. За си представим графично тази сложни обяснения, можем да представим т.н боксплот диаграма на фигура 3:



С помощта на интерквартилния размах можем да определим кои стойности са “екстремални” [16]

[15] Първи квартил се нарича още 25-ти перцентил - тоест 25% от пациентите са преди тази стойност

фигура 3: Можете да си представите боксплота като кутия с мустаци. Лявата страна на кутията представлява първият квартил - стойността е 2. Дясната страна на кутията представя 75-тия перцентил или 3-тия квартил (Q3) в случая това е стойността 4. В средата на кутията е представена медианата (стойността на втория квартил).

[16] Стойностите бегачи, могат да бъдат изчислени спрямо формулата на Тъки. Според нея всички стойности над $Q3 + 1.5 \cdot IQR$ или под $Q1 - 1.5 \cdot IQR$ са стойности бегачи или outliers. В нашия пример стойностите над $4 + 1.5 \cdot 2$ и под $2 - 1.5 \cdot 2$ са стойности бегълци.

Инферентна статистика

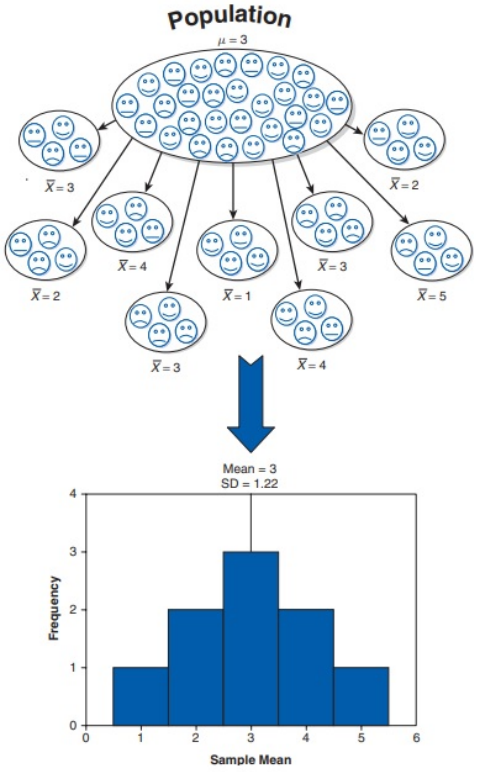
До момента описаните понятия са изцяло фокусирани върху извадката. Средната аритметична, например показва централната тенденция в извадката. Стандартното отклонение показва вариабилността на данните отново в извадката. Разполагайки с тези данни можем да направим извод за генералната съвкупност. Това е силата на статистическата теория. В основа на наблюдаваните студенти, можем да направим заключение за целия курс. Базисното условие, което приемаме, е че изучаваната извадка е случайна. Показателите средна аритметична, стандартно отклонение, медиана и интеквартилния размах изучавани в извадката се наричат статистики. Средната аритметична на пулса за една студентска група е статистика. В генералната съвкупност тези величини се наричат параметри.

> **!** Статистика и параметър
>
> Статистиката изчислена в извадка, използваме за да установим границите на параметъра в генералната съвкупност

Стандартна грешка на средната аритметична (SEM)

Можем ли да кажем, че средната стойност за извадката е просто равна на тази на генералната съвкупност? Нека си представим, че имаме една генерална съвкупност със стойност на средната аритметична точно равна на 3. Нека да направим 9 случайни извадки, от тази генерална съвкупност. Във всяка една от тях можем да определим средната аритметична. Ако нанесем тези 9 средни аритметични на една графика ще получим разпределние на средните аритметични което е нормално.

Колкото повече извадки правим от тази генерална съвкупност, средната аритметична от всички тези средни в самите извадки, ще е най-близо до истинската средна на генералната съвкупност. Това всъщност представлява закона на големите числа. В основа на това, ние можем да предположим


Population
$\mu = 3$
$\bar{X} = 3$
$\bar{X} = 2$
$\bar{X} = 4$
$\bar{X} = 3$
$\bar{X} = 2$
$\bar{X} = 1$
$\bar{X} = 5$
$\bar{X} = 3$
$\bar{X} = 4$
Mean = 3
SD = 1.22
Frequency
Sample Mean
0
1
2
3
4
5
6

с известна несигурност в какъв диапазон ще се намира средната аритметична в коя да е извадка от генералната съвпупност.

За да направим това е необходимо да изчислим **стандартната грешка**. Стандартната грешка е свързана с факта, че работим с извадки. Тя представлява точността на нашата точкова оценка "статистика" спрямо параметъра на генералната съвкупност.

За да изчислим, колко е точна средната аритметична за пулса на изследваната група спрямо средна аритметична за пулса на целия курс изпозлваме формулата за стандартна грешка [17]

Стандартната грешка в примера ни е 2,07. Сега можем да дадем някакъв интервал в които ще се намира средната сърдечна честота на целия курс (генералната съвкупност) в основа на средната аритметична и на нейната грешка. Този интервал се нарича интервал на доверителност

Интервал на доверителност средна аритметична

Интервалът на доверителност използваме за да ограничим, от колко - до колко може да бъде стойността на истинския параметър (средната честота на целия курс). интервалът на доверителност е базиран на една случайна извадка от цялата генерална съвкупност. Ние представяме този интервал с определена вероятност. **Например**, 95% , 99% или 90%.

Когато говорим за интервал на доверителност, можете да си представите следната графика.

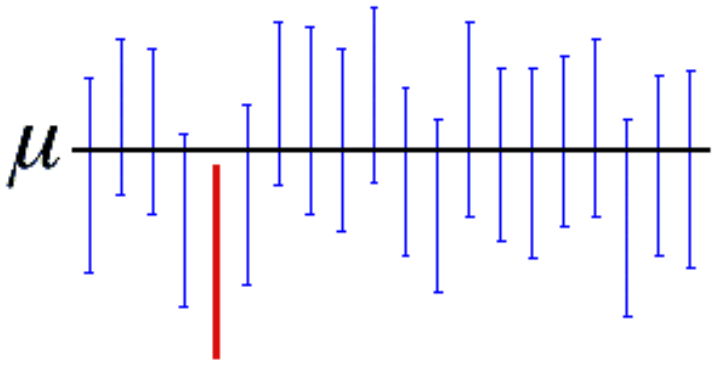
[17]

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

- Грешката зависи от стандартното отклонение- колкото по-голямо е то, толкова по-неточна е оценката.Грешката зависи и от обема на извадката n. Колкото повече студенти сме включили толкова по-малка ще е грешката на средната ни аритметична. Когато заместим във формулата получаваме:

$$SE_m = \frac{SD}{\sqrt{n}} = \frac{6.53}{\sqrt{10}} = \frac{6.53}{3.16} = 2.07$$

Така получихме стандартната грешка.



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.

Хоризонталната черна линия представя средната честотата на курса. Тази честота, която не знаем, но искаме да оценим. В случая сме направили 20 извадки от този курс- представете си, че сме измерили пулса на 20 случайни извадки от 5 студента. За всяка една група сме изчислили средната аритметична, стандартното отклонение и стандартната грешка. В основа на това сме направили интервал от две числа, между които предполагаме, че се намира истинската средна на целия курс. Този интервал ще е правилен в 95% от случаите, тоест е възможно от 20 извадки една (5%) ще представи интервал, които не съдържа истинската стойност на средната сърдечна честота на целия курс.

За да констриураме интерала на доверителност използваме формулата [18]. За да намерим интервала трябва първо да изберем неговата ширина. В медицината, най-често избираме **95% интервал на доверителност**. За да го изчислим трябва да знаем средната аритметична на извадката, грешката на средната аритметична и числото z. z се нарича **гаранционен множител** и зависи от това какъв интервал сме избрали. За 95% интервал на доверителност стойността на z е 1.95. Нека изчислим интервала на доверителност, за нашият пример с пулса [19].

[18]
$$CI = \bar{x} \pm z \cdot SE_m$$

[19]
$$95\% CI = 80.8 \pm 2.7 \cdot 1.96 = 80.8 \pm 5.29$$

Накрая можем да направим и заключение. В основа на нашата извадка, можем да кажем, че средната сърдечна честота на

курса в 95% от случайте ще бъде в диапазона 76 до 86 удъра в минута. Този интервал ще съдържа средната аритметична в 95% от всички извадки, които вземем от генералната съвкупност.

Стандартна грешка на процент

Стандартната грешка се изчислява не само за количествени променливи, но и за качествени такива. Нека вземем параметъра относителен дял на момичетата в 2-ри курс. Да речем, че нямаме достъп до целия курс, нямаме списъците на имената и не можем да разберем като изследваме всички. Затова можем да направим заключение в основата на една група - случайна извадка. Да приемем, че групата има 12 студента и 6 от тях са момичета. Това означава, че нашата точкова оценка относителен дял на момичетата в извадката е 50%.

Първо, подобно на средната аритметична е необходимо да изчислим каква е точността на тази оценка. За да ичзислим точността на този относителен дял ползваме формулата за стандартна грешка за относителен дял [20]

[20]

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

В нашия пример

$$SE_p = \sqrt{\frac{\hat{p} \cdot (100 - \hat{p})}{n}} = \sqrt{\frac{50 \cdot 50}{12}} = \sqrt{\frac{2500}{12}}$$

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

Интервал на доверителност за пропорция

За да изчислим интервала на доверителност за относителен дял отново изпозлваме формулата за CI [21]. Стойността на z избираме, спрямо това какъв доверителен интервал искаме да използваме. Понеже в медицинските изследвания се използва 95% интервал ще използваме стойността на z= 1.96 [22].

Извода е подобен - ако вземем 100 извадки от тази генерална съвкупност (всички студенти 2 курс) процента на момичета в 95% от случаите, ще бъде между 21.8% и 78.2%. Този интервал е доста широк. Причината за това е имаме много малко наблюдения. С увеличаването на броя на наблюденията грешката намалява, а интервала се скъсява

[21]

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

[22]

$$95\% CI = \hat{p} \pm z \cdot SE_p = 50\% \pm 1.96 \cdot 14.4\%$$

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

Можете да намерите всички формули от днешното занятие тук https://www.raredis.work/edu/Formulas%20and%20tables.pdf

Важно!

Ако изберем да използваме 99% интервал на доверителност, ще използваме по-висока стойност на гаранционния множител z= 2.57 - това разбира се ще доведе до по-широк интервал. Тоест, ако искаме нашият интервал да е валиден за повече случайни извадки от генералната съвкупност, то трябва да го разширим. Увеличавайки интервалът на доверителност, възможността нашият интервал да пропусне истинската стойност ще намалее, но това коства увеличаването на ширината му. Ако изберем да използваме 90% интервал на доверителност, тогава ще умножаваме по z=1.65. Това ще доведе до по-тесен интервал. От своя стана, обаче възможността да пропуснем истинката стойност на параметъра се увеличава до 10%. Тоест намаляйки интервала на доверителност, увеличаваме възможността да пропуснем истинската стойност в генералната съвкупност.