Приложна статистика

Базови понятия и Статистически методи

Contents

[1 Измервания и статистически методи 3](#_Toc171182668)

[1.1 Основни видове измерителни скали 3](#_Toc171182669)

[1.1.1 Качествени променливи 3](#_Toc171182670)

[1.1.2 Количествени променливи 4](#_Toc171182671)

[1.2 Приложение на статистическите методи 4](#_Toc171182672)

[2 Честотен анализ 5](#_Toc171182673)

[2.1 Обща идея за задачите 5](#_Toc171182674)

[2.2 Едномерно разпределение 5](#_Toc171182675)

[2.3 Двумерно разпределение 6](#_Toc171182676)

[3 Вариационен анализ. Разпределения 6](#_Toc171182677)

[3.1 Вариация 6](#_Toc171182678)

[3.2 Показатели за средно равнище 7](#_Toc171182679)

[3.3 Показатели за разсейване 11](#_Toc171182680)

[3.3.1 Пример за коефициент на вариация 14](#_Toc171182681)

[3.4 Нормално разпределение 14](#_Toc171182682)

[4 Взаимно свързани зависимости - Корелации 19](#_Toc171182683)

[4.1 Зависимости 19](#_Toc171182684)

[4.2 Корелационен анализ 20](#_Toc171182685)

[4.2.1 Коефициент на обикновена линейна корелация на Пирсън (r) 22](#_Toc171182686)

[4.2.2 Коефициент на рангова корелация на Спирман () 23](#_Toc171182687)

[4.3 Статистическа значимост на коефициентите на корелация 23](#_Toc171182688)

[4.4 Крос-корелация (примери с 2 сигнала) 24](#_Toc171182689)

[5 Параметрични и непараметрични критерии 24](#_Toc171182690)

[5.1 Фактори, изследвани с параметрични и непараметрични критерии. 24](#_Toc171182691)

[5.2 Дали повечето случайни променливи (величини) са нормално разпределени? 25](#_Toc171182692)

[5.3 Размерът на извадката 25](#_Toc171182693)

[5.4 Проблемът с измерванията 25](#_Toc171182694)

[5.5 Параметрични и непараметрични методи 25](#_Toc171182695)

[6 Проверка на хипотези при независими извадки 27](#_Toc171182696)

[6.1 Хипотеза и алтернатива 27](#_Toc171182697)

[6.2 Поредица от стъпки за проверка на хипотези 28](#_Toc171182698)

[6.3 Критерии за проверка на хипотези 29](#_Toc171182699)

[6.4 Сравняване на средно равнище на признака при две независими извадки. t -критерий на Стюдънт за независими извадки 30](#_Toc171182700)

[6.5 Сравняване на средно равнище на признака на три и повече независими извадки. F-критерий на Фишер (еднофакторен дисперсионен анализ) 31](#_Toc171182701)

[7 Зависими извадки 32](#_Toc171182702)

[7.1 Показатели за описание на развитие 32](#_Toc171182703)

[7.2 Сравняване на средни величини при зависими извадки (t-критерий на Стюдънт за зависими извадки) 32](#_Toc171182704)

[7.3 Приложение на проверката на хипотези при обработка на данни от педагогически експеримент 33](#_Toc171182705)

# Измервания и статистически методи

## Основни видове измерителни скали

Статистическите единици (изследваните лица) се характеризират със свои белези, качества, свойства, по които си приличат или се различават. Тези техни характеристики се наричат признаци. В процеса на измерването на признаците се присвояват числа. Тъй като показателите, които описват поведението и състоянието на обектите варират – т.е. при отделните изследвани лица заемат различни стойности – те представляват променливи величини. Изключително важни моменти, с които трябва да бъде съобразено въвеждането на данните, а след това обработката и анализът им са:

* Естеството на признаците и по-специално на какво (**количествено или качествено**) описание се поддават.
* По какъв начин е измерен признакът в конкретното изследване, т.е. **в каква измерителна скала е изразен**.

В зависимост от естеството на признаците и измерителната скала, в която са изразени, променливите величини се делят на две големи групи - качествени и количествени.

### Качествени променливи

**Качествени** (неметрични) са променливите, които характеризират качествени особености на обектите. При тях се описва категорията, към която принадлежи обектът или **интензитета на притежавания признак**, без да има количествена информация за него. Те биват 3 вида:

* **Номинално скалирани (категорийни)** са променливите, при които обектите са разпределени в еднородни равностойни групи (например вид изучавана дисциплина). Ако признакът има само две възможни състояния (например пол), той се нарича **алтернативен**. **При въвеждане на качествени променливи съответните категории се записват с числа (кодове), които обаче не носят количествена информация за обектите, а само показват принадлежността им към дадена качествена група**. При дефиниране на качествени променливи в *SPSS* се задават кодът и наименованието на съответната категория.
* **Ординално скалирани** са променливите, при които **обектите се разпределят в еднородни групи, но между тях съществува известно степенуване на интензитета на притежавания признак** (напр. по възраст - "млади", "средна възраст", "стари"; по доход - "нисък", "среден", "висок" и други). Тези променливи се дефинират както номиналните, но при обработката могат да се ползват по-мощни статистически методи, които да отчитат различията в интензитета на притежавания от отделните групи признак.
* Рангово скалирани са променливите, при които обектите се подреждат (ранжират) въз основа на определени правила. Мястото на даден обект в ранговата редица се нарича рангов номер. Ранжирането се прави по различен начин – въз основа на експертни оценки, регламент на дадено състезание и др. За обработка на рангово скалирани признаци се ползват методите на корелационния анализ.

### Количествени променливи

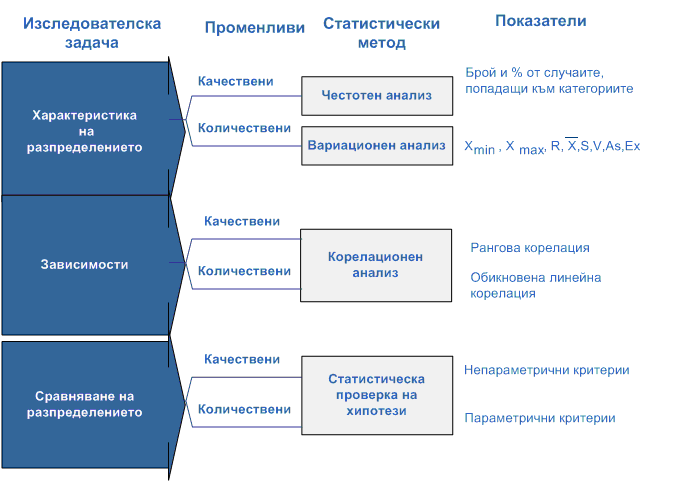
**Количествените (метрично скалирани)** променливи носят информация за количеството на притежавания признак и са изразени в съответни мерни единици (*m, s, kg, брой*). Те се делят на две групи:

* Интервално скалирани, при които се отчита количеството на притежавания признак, но значение 0 не означава отсъствието на признака, а е някаква условно възприета стойност (например температура, измерена по скалата на Целзий).
* Пропорционално скалирани, при които нулевото значение на признака е абсолютно (например 0о по Келвин).

При изброяване на видовете променливи величини от гледна точка на измерителната скала, в която са изразени, **е спазена последователността за повишаване способността на дадена скала да отчита различията в притежавания признак, т.е. нейната сила**. *Количествените измерителни скали са “силни” скали, а качествените – “слаби”*. Възможно е количествен признак да бъде измерен и качествено, но обратното – преминаване от “слаба” към “силна” скала, не е възможно!!!!!!!!!!

## Приложение на статистическите методи

На следващата схема е изобразено всичко (т.е. принципите) на статистическите изследвания и следователно следвайки тази схема може да се отговори на много въпроси, свързани с принципите на статистиката и приложението на нейните методи при решаване на изследователски задачи.



Следвайки горната схема, започваме да описваме методите един след друг, като тръгваме от горе на долу. По този начин ще отговорим на всеки конкретен въпрос в дълбочина.

# Честотен анализ

## Обща идея за задачите

Да се характеризира разпределението на категорийна променлива означава **да се опишат възможните състояния (категории) на променливата и съответстващата им абсолютна или относителна честота.**

При обработка на категорийни променливи биха могли да се решават следните изследователски задачи:

* Да се опише разпределението на една категорийна променлива (т. нар. едномерно разпределение).
* Да се опише разпределението на два или повече свързани категорийни променливи (т.нар. двумерно и многомерно разпределение).

## Едномерно разпределение

**Да се характеризира разпределението** на номинално или ординарно скалирана променлива означава да се опише колко статистически единици принадлежат към всяка категория.

В таблица 2.1 са представени резултати от анкетно проучване сред студентите (n=48), дипломирали се през 2004 г. Зададен е въпросът "Доволни ли сте от обучението?"

**Таблица 2.1.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Категории** | **Брой (f)** | **Процент (W)** | **Кумулативен процент (Wcum)** |
| Категорично да | 26 | 54,2 | 54,2 |
| По-скоро да | 15 | 31,2 | 85,4 |
| По-скоро не | 7 | 14,6 | 100,0 |
| **Общо** | **48** | **100,0** |  |

**Абсолютната честота (f)** показва колко статистически единици принадлежат към съответната категория (например 26 студенти са отговорили с "Категорично да" - f = 26).

**Относителната честота (W)** показва каква част (ако е умножена по 100 – какъв процент) от общия брой случаи принадлежат към съответната категория (студентите отговорили с "Категорично да" представляват 54,2% от изследваните). Изчислява се по формула:

* + е абсолютната честота; обем на извадката;

**Относителна кумулативна честота ()** описва **каква част *(%)* от общия брой случаи попадат в съответната и предходните категории, взети заедно** (например 85,4% от изследваните лица са отговорили с "Категорично да" и "По-скоро да").

## Двумерно разпределение

Ако към предходния пример се добави информация за образователната степен, която са завършили изследваните студенти, резултатите могат да се представят под формата на двумерно разпределение на променливите “Доволен ли сте от обучението?” и “Образователно квалификационна степен” , както е показано в таблица 2.2.

Таблица 2.2

|  |  |  |  |
| --- | --- | --- | --- |
| Отговори | Образователно квалификационна степен | | Общо |
| бакалавър | магистър |
| Категорично да | 15 | 11 | 26 |
| По-скоро да | 10 | 5 | 15 |
| По-скоро не | 5 | 2 | 7 |
| **Общо** | **30** | **18** | **48** |

Тези таблици се наричат **крос-таблици** (свързани редове). Статистическите показатели са същите, както описаните – абсолютна () и относителна честота ().

# Вариационен анализ. Разпределения

Това е най-математическата част на изложението, но ще трябва да се прочете и изучи!!!!!

## Вариация

Изследваните лица се различават (варират) по величината на изследваните признаци. Вариацията е присъщо свойство на статистическите съвкупности. Ако тя не съществуваше, статистическите проучвания щяха да бъдат излишни. Тези различия са породени от различни по своето естество фактори, които най-общо се разделят на две категории:

* **закономерни**, които действат еднакво върху всички изучавани единици и определят типичното, закономерното състояние на признака (например типично, характерно за баскетболистите е, че имат висок ръст).
* **случайни**, които действат с различна сила върху изучаваните единици и водят до индивидуални отклонения във величината на признака (отделните баскетболисти се различават по своя ръст).

Задача на статистическото проучване е да даде обобщаващи количествени характеристики на състоянието на изследвания признак. За целта е необходимо, **на първо място**, да се установи какво е типичното за дадената съвкупност значение на признака, за да се опише влиянието на закономерно действащите фактори. **На второ място**, от особена важност е характеризиране на разсейването, варирането на признаците, за да се отчете влиянието на случайните фактори.

## Показатели за средно равнище

Показателите за средно равнище представляват обобщаващи количествени характеристики, с помощта на които се описва типичното (характерното) състояние на изследвания признак. Най-често употребяваните показатели са **мода (), медиана () и средна аритметична величина ().**

* **Модата** **()** е стойността, която се повтаря най-голям брой пъти. . (**зависи само от обем**)
* **Медианата ()** е онази стойност на променливата величина, която заема средно положение във вариационния ред и го дели на две равни части. Броят на стойностите, които са по-малки или равни на медианата, е равен на броя на тези, които са по-големи или съвпадат с нея. (**зависи само от стойности и не разглежда обемите** им)

За всички следващи дефиниции ще използваме: означенията: всяка конкретна стойност в извадката; обем на извадката;

* **Средно аритметично** **()**:
* *сумирани всички стойности;*
* **Средно геометрично** **()**
* *произведение на всички стойности;*
* **Средна хармонична стойност** **()**
* *сума от реципточните на всички стойности;*
* **Средно квадратично** **()**:
* *всяка конкретна стойност в извадката;*
* *сумирани квадратите на всички стойности;*
* *обем на извадката;*

Сега ще коментираме показателите за средно равнище в малко по-широк контекст. Нека започнем с едно известно неравенство между средните стойности, математическото доказателство на което не е тривиално, но няма да се спираме на него, а просто ще запомним и използваме неравенството.

**Пример:** *Какъв е смисълът на различните средни стойности?*

**Сравнение на Средно претеглена аритметична и Средно претеглена хармонична стойност!**

Средно-хармонична стойност **()** е средна стойност, подходяща за ситуации, когато се търси средна ставка. **Опростява изчисленията и многократно подобрява резултата. Няма нужда от коригиращи коефициенти!**

Средната хармонична стойност може да бъде изразена като реципрочна на средната аритметична на реципрочните стойности на дадения набор от наблюдения (дефиницията по-горе): , където е всяка стойност в извадката; обем на извадката; сума от реципточните на всички стойности;

**Свойства:**

Хармоничната средна е вдлъбната по Щур функция и доминирана от минимума от нейните аргументи, в смисъл, че за всеки положителен набор от аргументи, , . По този начин средната хармонична стойност не може да бъде произволно голяма чрез промяна на някои стойности към по-големи (докато поне една стойност остава непроменена).

Хармоничната средна е вдлъбната, което е по-силно свойство от вдлъбнатостта на Щур. Трябва да се ползват само положителни числа, тъй като средната стойност не е вдлъбната, при отрицателни стойности.

**Илюстрация**

Средно претеглената средна хармонична стойност е предпочитаният метод за осредняване на произведения, като например съотношението ***цена-обем*** или скорости ***време-път***. Ако тези съотношения са осреднени с помощта на ***средно претеглена аритметична стойност***, на високите точки от данни се дават по-големи тегла, отколкото на ниските точки от данни. Нещо повече, средно претеглената хармонична стойност осигурява правилната средна стойност за разлика от средно претеглената аритметична стойност.

***Да разгледаме следния пример:*** Превозно средство изминава разстояние от в едната посока със скорост и се връща със скорост . Въпросът е колко е средната му скорост?

Очевидно превозното средство се е движило общо 4 часа (1 час на отиване и 3 часа на връщане) и е изминало общо . Тогава средната му скорост е:

Нека не знаем пътя а разполагаме само със съотношенията ***време-път*** , т.е. със скоростите.

Да видим какъв резултат дава **средно аритметичното** на скоростите:

Нека сега изчислим **средно хармонично** на скоростите:

Следователно реалната средна скорост е средната хармонична стойност на скоростите (), а не средната аритметична ().

**Претеглената хармонична средна стойност, претегля правилно всяка точка от данни.** [1]

Когато се прилага към нормализирани съотношения, средно аритметичната претеглена стойност, е отклонена нагоре и не може да бъде числено обоснована, тъй като се основава на изравнени печалби; точно както скоростите на превозните средства не могат да бъдат осреднени за двупосочно пътуване.[2]

**Други примери за използване на показателите за средно равнище**

**Пример-0** за използване на показателите за средно равнище, базиран на симулирани данни!

Нека са дадени следните 10 цени между 0 и 9: 4 7 6 3 1 6 1 3 3 5

Нека първо да ги подредим по големина: 1 1 3 3 3 4 5 6 6 7

**Модата** ***Mo*** не се интересува от конкретните стойности на цените, а само от техния обем. С други думи от ***Mo*** е равно на онази цена, която е представена най-много пъти в редицата от цени. За да изчислим модата ще намерим максимума на хистограмата на цените **фиг. 3.2.A)**. т.е. ***Mo*** = 3. Ако има няколко цени с еднакъв (най-голям) обем срещания, то ***Mo*** е равно на средното аритметично на цените с най-голям обем.

**Медианата** за разлика от модата не се интересува от обема, а само oт различните цени, подредени по големина. В случая различните цени са: 1 3 4 5 6 7. Медианата е равна на онази цена, която се намира в центъра на редицата уникални цени. В случая това са две цени цена 3 и цена 4. Ако броят на цените е четем, то ***Me*** = Средното аритметично на средните две цени, ***Me*** = 3.5, **фиг. 3.2.B).**

|  |  |
| --- | --- |
| A) | B) |
| C) | |

**фиг. 3.2.**

Когато трябва да се работи със средно-претеглена цена, трябва да се вземат под внимание едновременно и стойността на цената и нейния обем. Тогава използваме някое от описаните по-горе четири средни **фиг. 3.2.C).**

## Показатели за разсейване

Показателите за разсейване дават количествена характеристика на различията на стойностите. За измерване на статистическото разсейване най-често се използват показателите **размах (), стандартно отклонение (), коефициент на вариация ().**

* **Размахът “range” ()** е най-елементарният измерител на статистическото разсейване. Той представлява разликата между най-голямата и най-малката стойност на променливата.

*Тъй като в изчисляването на размаха участват само двете крайни стойности на вариационния ред, той е твърде* ***неустойчив измерител на разсейването!!!!***

* **Стандартното отклонение** “***standard deviation***“ () е най-прецизният и често използван показател за разсейване. Той описва степента на отклонение на стойностите на променливата величина от средната аритметична. Формулата, която го дефинира е:

Това което написахме за стандартното отклонение (***standard deviation***) е доста странно и се нуждае от повече обяснения, за да се изясни как величините преминават една в друга, как се изчисляват и кои характеристики на извадката се запазват, засилват или изгубват. Това разбиране е съществено за по-нататъшното им използване. Какво казва логиката? Имаме:

* обем на извадката, следователно това е броят опити, които сме направили и в крайна сметка разполагаме с числа;
* всяка конкретна стойност в извадката, значи по този начин означаваме всяко от числата ;
* сумирани всички стойности;
* средното (***mean***), или сумата на числата, делени на броя им;
* на всяка една стойност по отношение на средната, или получаваме ;
* забелязваме следната неприятност: едни от числата ще са положителни, а други – отрицателни!!! Как да се справим с това? Естествено вдигаме на квадрат и всички членове са положителни!
* получаваме което представлява една нова редица от числа, показващи отклоненията на измерванията от средната стойност (разбира се на квадрат). ***Нашето вероломство няма граници и затова ще се опитаме да намерим средното на тази нова редица***, т.е. за нея да приложим същия подход, както за началните данни… разбира се накрая ще коренуваме, за да се върнем в същата скала.
* Изчисляваме и делим на броя им;
* така, че получихме ново средно (средно на отклоненията от средното) и остава да коренуваме и ще получим бленуваната формула;
* Това се нарича стандартно отклонение или стандартна девиация! А тогава какво беше онова, което написахме преди? Има разлика в знаменателите! Сега пишем , а преди написахме Човек трябва да мисли и да обръща внимание на тези неща!!!!!!!

Къде се корени проблемът? Формулата с е валидна за всички опити (Генералната съвкупност), т.е. там се предполага граничен преход към безкрайност, а формулата с е валидна за извадката. Това се нарича поправка на ***Bessel*** – повече няма да задълбавам в тази посока, но може да се зададе въпрос и трябва да се отговори.

Ето какво казва ***Wikipedia*** по въпроса + хубав пример:

***Standard deviation*** *is a widely used measurement of variability or diversity in* [*statistics*](http://en.wikipedia.org/wiki/Statistics) *and* [*probability theory*](http://en.wikipedia.org/wiki/Probability_theory)*. It shows how much variation or "*[*dispersion*](http://en.wikipedia.org/wiki/Statistical_dispersion)*" defers from the average (*[*mean*](http://en.wikipedia.org/wiki/Mean) *or expected value). For example, a low standard deviation indicates that the data points are very close to the* [*mean*](http://en.wikipedia.org/wiki/Mean)*. In contrast, a high standard deviation indicates that the data are spread out over an extensive range of values.*

*Technically, the standard deviation of a* [*statistical population*](http://en.wikipedia.org/wiki/Statistical_population)*, data set, or* [*probability distribution*](http://en.wikipedia.org/wiki/Probability_distribution) *is the* [*square root*](http://en.wikipedia.org/wiki/Square_root) *of its* [*variance*](http://en.wikipedia.org/wiki/Variance)*. It is* [*algebraically*](http://en.wikipedia.org/wiki/Algebra) *simpler though practically less* [*robust*](http://en.wikipedia.org/wiki/Robust_statistics) *than the* [*average absolute deviation*](http://en.wikipedia.org/wiki/Average_absolute_deviation)*.*[*[1]*](http://en.wikipedia.org/wiki/Standard_deviation#cite_note-0)[*[2]*](http://en.wikipedia.org/wiki/Standard_deviation#cite_note-1) *A useful property of standard deviation is that, unlike* [*variance*](http://en.wikipedia.org/wiki/Variance)*, it is expressed in the same units as the data.*

*In addition to expressing the variability of a population, standard deviation is commonly used to measure* [*confidence in statistical conclusions*](http://en.wikipedia.org/wiki/Confidence_in_statistical_conclusions)*. For example, the* [*margin of error*](http://en.wikipedia.org/wiki/Margin_of_error) *in* [*polling*](http://en.wikipedia.org/wiki/Opinion_poll) *data is determined by calculating the expected standard deviation in the results if the same poll were to be conducted multiple times. The reported margin of error is typically about twice the standard deviation ­– the radius of a 95 per cent* [*confidence interval*](http://en.wikipedia.org/wiki/Confidence_interval)*. In* [*science*](http://en.wikipedia.org/wiki/Science)*, researchers commonly report the standard deviation of experimental data, and only effects that fall far outside the range of standard deviation are considered* [*statistically significant*](http://en.wikipedia.org/wiki/Statistically_significant) *– normal random error or variation in the measurements is, in this way, distinguished from causal variation. Standard deviation is also substantial in* [*finance*](http://en.wikipedia.org/wiki/Finance)*, where the standard deviation on the* [*rate of return*](http://en.wikipedia.org/wiki/Rate_of_return) *on an* [*investment*](http://en.wikipedia.org/wiki/Investment) *is a measure of the* [*volatility*](http://en.wikipedia.org/wiki/Volatility_(finance)) *of the investment.*

*When only a* [*sample*](http://en.wikipedia.org/wiki/Statistical_sample) *of data from a population is available, the population standard deviation can be estimated by a modified quantity called the sample standard deviation.*

**Пример (*Wikipedia*):**

*A slightly more complicated real-life example is that the* [*average height for adult men*](http://en.wikipedia.org/wiki/Human_height#Average_height_around_the_world) *in the* [*United States*](http://en.wikipedia.org/wiki/United_States) *is about 70", with a standard deviation of around 3". It means that most men (about 68%, assuming a* [*normal distribution*](http://en.wikipedia.org/wiki/Normal_distribution)*) have a height within 3" of the mean (67"–73") — one standard deviation — and almost all men (about 95%) have a height within 6" of the mean (64"–76") — two standard deviations. If the standard deviation were zero, all men would be exactly 70" tall. If the standard deviation were 20", then men would have much more variable heights, with a typical range of about 50"–90". Three standard deviations account for 99.7% of the sample population being studied, assuming the distribution is normal (bell-shaped).*

Преди да продължим, ми идва ***една еретична мисъл***, и аз ***не бих могъл да устоя на изкушението не да въведа ред и термини в тези величини***! Нека означим:

Тогава дефинираме тези изрази, като моменти (централни) на изследваната величина – съответно от първи до -ти, като за нулев момент ще считаме средното!!!

Изброените статистически показатели за разсейване изразяват степента на различията в същата мерна единица, в която е изследваният признак. Това не дава възможност за сравняване на вариацията на признаци, изразени в различни мерни единици, а в случай че са в една и съща мерна единица – относно различно средно равнище. За разрешаване на проблеми от подобно естество в статистиката се използва **коефициентът на вариация ()**. Той се изчислява по формула:

Където:

* - стандартното отклонение
* - средната стойност
* **Коефициентът на вариация** () дава информация за разсейването на признака, изразено в проценти, което дава възможност за сравняване на вариацията на различни признаци. Освен това се ползва за оценяване на еднородността на извадката:
* Счита се, че разсейването на признака е малко (**извадката е еднородна**), когато стойността му е до 10-12%.
* Между 10 и 30% извадката е **приблизително еднородна**.
* Когато е над 30% разсейването на признака е голямо (**извадката е силно нееднородна**).

### Пример за коефициент на вариация

*Фред* иска да намери нова инвестиция за портфолиото си. Той търси безопасна инвестиция, която осигурява стабилна възвръщаемост. Той разглежда следните възможности за инвестиция:

* **Акции: На** *Фред* бяха предложени акции на **ABC Corp**. Това е зряла компания със силни оперативни и финансови резултати. Волатилността на акциите е 10%, а очакваната възвръщаемост е 14%.
* **ETF:** Друг вариант е борсово търгуван фонд (ETF), е популярно средство за инвестиране, при което портфейлите могат да бъдат по-гъвкави и диверсифицирани в широк спектър от всички налични класове активи. ETF предлага очаквана възвръщаемост от 13% при волатилност от 7%.
* **Облигации:** Облигациите с отличен кредитен рейтинг предлагат очаквана възвръщаемост от 3% с 2% волатилност.

За да избере най-подходящата възможност за инвестиция, *Фред* реши да изчисли коефициента на вариация на всяка опция. Използвайки горната формула, той получи следните резултати:

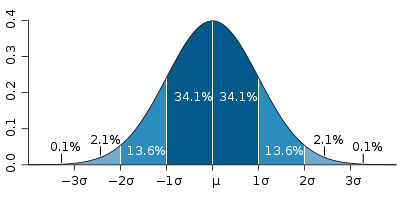
Въз основа на изчисленията по-горе, *Фред* ще инвестира в ETF, тъй като той предлага най-ниския коефициент (вариация) с оптималното съотношение риск-печалба.

## Нормално разпределение

В настоящото изложение многократно е използван терминът “разпределение на признака”, с което се означава съответствието между стойностите на променливата величина и тяхната абсолютна или относителна честота. Тъй като статистическите проучвания се основават на изследването на извадки, това разпределение се нарича **емпирично.** Ако то се онагледи графично, по конфигурацията на графичния образ може да се съди за закономерността, на която е подчинено изучаваната променлива величина, и да се търси аналог сред т. нар. **теоретични разпределения**.

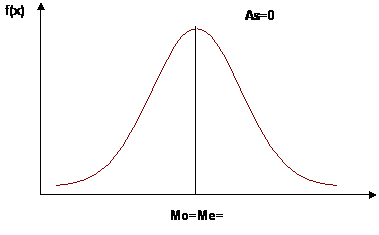
Изключително важно място в теорията на статистиката заема **нормалното разпределение**. То носи наименованието **Гаус-Лапласово** по имената на немския математик **Карл Гаус** (1777 – 1855) и френския математик **Пиер Лаплас** (1739 – 1827), които са го изследвали и описали теоретично. Нормалното разпределение стои в основата на изясняване на важни положения от репрезентативните статистически проучвания, интервалното оценяване, проверката на хипотези, методите за изготвяне на нормативи и др. Ще се спрем накратко на някои негови особености:

Характерна за нормалното разпределение е **камбановидната крива на разпределение** (фигура 3.1)- стойностите около средното равнище на признака имат най-голяма честота. Колкото повече се отдалечават от центъра на разпределението – толкова по- рядко се срещат.



Горната фигура не е номерирана, тъй като на нея са нанесени повечето от данните и свойствата, обсъждани по-надолу. Тя важна и критериална крива и не трябва да се забравя!

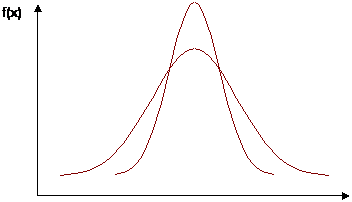
Сега ще коментираме разпределението подробно на части.



*Фигура 3.1. Крива на нормалното разпределение*

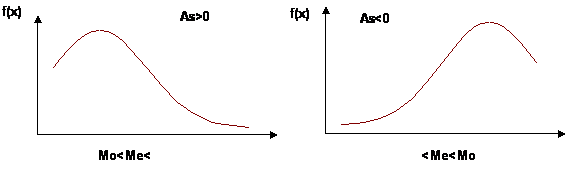
**Средната аритметична величина** , **модата** и **медианата** са равни и се намират в центъра разпределението.

**Стандартното отклонение () определя заостреността** **на кривата на разпределение**. Колкото отклоненията от средното равнище са по-малки, толкова кривата е по-изострена и обратно. На фиг. 3.2 са представени две емпирични разпределения с еднаква форма и център, но с различно разсейване.



*Фигура 3.2. Крива на разпределението при различно разсейване на признака*

Кривата на нормалното разпределение е симетрична по отношение на перпендикуляра, спуснат от нейния връх към абсцисата, където се намира средната стойност. На фигура 3.3 са представени 2 несиметрични разпределения. Разпределение (а) е с дясно изтеглено, а (b) - с ляво изтеглено рамо.



*Фигура 3.3. Симетричност на разпределението*

Симетричността на разпределението се описва с показателя за **асиметрия** (), който при нормално разпределение е равен на нула (). Когато коефициентът на асиметрия е с положителен знак, кривата на разпределение е с дясно изтеглено рамо (**а**), средната аритметична е по-голяма от модата и медианата. При разпределение с ляво изтеглено рамо (**b**) стойността на коефициента на асиметрия е с отрицателен знак, а средната аритметична е по-малка от модата и медианата.

За симетричността на емпиричното разпределение се съди по изчисления от извадката по формула 3.5 **емпиричен** **коефициент на асиметрия** ().

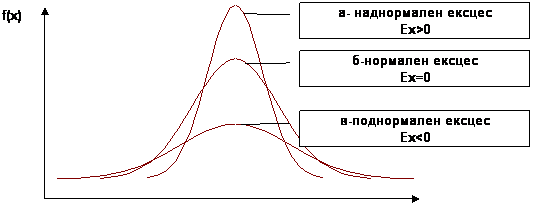
Има **таблични стойности на критерия** () в зависимост от обема на извадката () и равнището на значимост **().** В случай че:

се приема, че разпределението на емпиричните данни е симетрично.

се счита, че разпределението е несиметрично.

**Височината на върха на нормалното разпределение се описва с показателя ексцес (),** който при нормално разпределение е равен на нула.

На фигура 3.4 са представени примерни разпределения с нормален (**b**), висок - положителен (**а**) и нисък - отрицателен ексцес (**с**).



*Фигура 3.4. Разпределения с различна височина на върха*

Информация за степента на съвпадение на височината на върха на емпиричното разпределение с приетия за еталон връх на нормалното разпределение () дава изчисленият по формулата емпиричен коефициент на ексцес ().

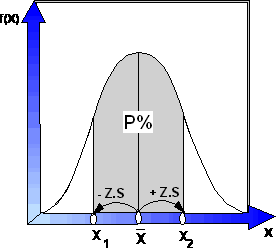
Табличните стойности на коефициента на ексцес () в зависимост от обема на извадката () и равнището на значимост **()**. В случай че:

се приема, че разпределението на емпиричните данни има нормален ексцес.

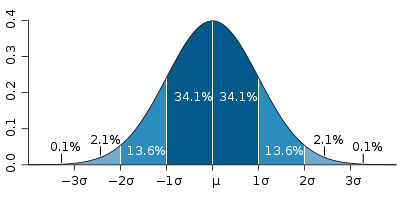
се счита, че разпределението има ненормален ексцес.

Площта, заключена между нормалната крива, абсцисата и перпендикулярите, спуснати към тази ос в точки

представлява вероятността (), дадена стойност на случайната величина да се намира в интервала от до . Вероятността не зависи от стойността на средната аритметична и стандартното отклонение, а само от множителя

[](http://basaga.org/wiki/index.php?title=%D0%A4%D0%B0%D0%B9%D0%BB:Variacionenanaliz16.png)

Дадени са стойностите на и процентът от случаите (), който попада в съответния интервал.



Това важно свойство на нормалното разпределение се нарича “**правило на трите сигми**”.

От фигурата се вижда, че в интервала:

* От до попадат централните 68.23% от случаите ().
* От до попадат централните 95.44% от случаите ().
* От до попадат централните 99.74% от случаите (). Вероятността дадена стойност на променливата величина да е извън този интервал е 0,25%, което от практическа гледна точка означава, че ако даден признак има нормално разпределение, всички значения на променливата попадат в границите на .

Свойствата на нормалното разпределение имат голямо приложение в научно-изследователската дейност. Те дават възможност да се оценява всяко едно значение на променливата величина. За целта се ползва т.нар. -оценка (), която представлява нормираната стойност на оценяваното значение. Очевидно е, че ако стойността на оценяваната величина е под средното равнище, нейната –оценка е отрицателна. Ако съвпада с центъра на разпределението, е 0, а ако е по-голяма от него - положително число.

# Взаимно свързани зависимости - Корелации

## Зависимости

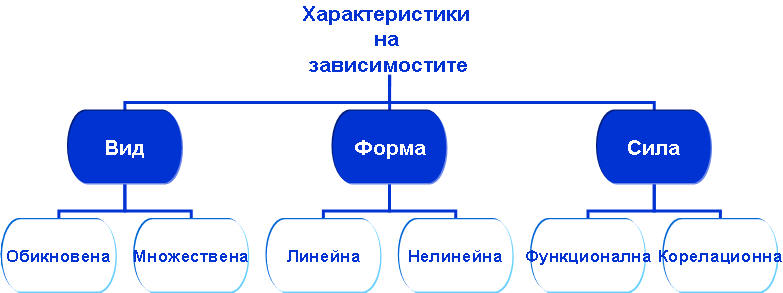
Статистиката изучава зависимостите, проявяващи се в масовите явления. При тях връзката между явлението “*фактор*” и явлението “*резултат*” се съдържа във всеки отделен случай, но може да бъде доловена само на базата на масово наблюдение. Единствено тогава се елиминира влиянието на случайните фактори и се проявява закономерната връзка между изучаваните явления.

Целта на голяма част от научните изследвания се състои в разкриване на факторите на дадено явление, изграждане на модели за оптимизирането и прогнозирането му. За решаване на проблеми от подобен характер се използват статистическите методи за изследване на зависимости.

Понятието **зависимост** означава връзка между две или повече променливи величини. Когато зависимостта е изразена математически, тя се нарича функция. Най-общият аналитичен вид на една зависимост е: , където и са променливите величини:

* се нарича аргумент на функцията или **независима променлива**.
* е **зависима променлива** (тъй като се предполага, че зависи от ) или е функция на .

Основните характеристики на зависимостите са вид, форма и сила - фиг. 4.1.



**Фиг. 4.1.**

**Зависимостите са:**

**В зависимост от броя на факторите, които се изследват**:

* **Обикновена** - Изследва се влиянието на един фактор: .
* **Множествена** - Изследва се влиянието на два или повече фактора .

**В зависимост от математическия и графичен модел на зависимостта:**

* **Линейна** - Промяната на фактора () води до пропорционална промяна на функцията (). Графиката е права линия.
* **Криволинейна** - Промяната на фактора (), води до непропорционална промяна в . Графиката е крива линия.

**В зависимост от степента на съвпадение на фактическите стойности на Y с теоретичните**

* **Функционална** - Пълна зависимост – в регресионния модел участват всички фактори, влияещи върху .
* **Корелационна** - Непълна зависимост - в регресионния модел не участват всички фактори, влияещи върху , а само част от тях.

Разграничаването на зависимостите по тези характеристики е важно, тъй като подходите за анализ, статистическите показатели, които се изчисляват и интерпретират, при всяка една от тези разновидности са различни.

Когато се проучват зависимости, обикновено се поставят две основни познавателни задачи:

1. Да се изследва формата на зависимостта и се опише математически връзката между променливите величини (т.е. да се изчисли аналитичния вид на функцията). Тази задача се решава посредством **регресионния анализ**.
2. Да се установи силата на зависимостта, за което се използва **корелационния анализ**

## Корелационен анализ

Състоянието или поведението на човека в условията на спорта, се формира под въздействието на множество фактори. Като се изследва влиянието само на един или няколко от тях не е възможно да се предвиди точно състоянието на явлението “резултат”. Поради това зависимостите са непълни или корелационни.

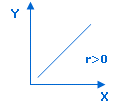
Основната задача на корелационния анализ е да опише количествено силата (теснотата) на корелационните зависимости. **Статистическите показатели, които носят тази информация, се наричат коефициенти на корелация**. Те заемат стойности от –1 до +1.

Абсолютната стойност на коефициента носи информация за силата (степента) на зависимост между Х и Y, както е описано в таблица 4.1.

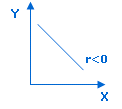
Таблица 4.1

|  |  |
| --- | --- |
| **Стойност на | r |** | **Сила (степен) на зависимост** |
| **r=0** | Липсва зависимост |
| **До 0,3** | Слаба |
| **От 0,3 до 0,5** | Умерена |
| **От 0,5 до 0,7** | Значителна |
| **От 0,7 до 0,9** | Голяма |
| **Над 0,9** | Много голяма |
| **r=1** | Функционална зависимост |

В случай че става дума за линейна зависимост, на интерпретация се подлага и знакът на **коефициента на корелация** – той носи информация за посоката на зависимостта:



При зависимостта е възходяща (еднопосочна) - на по-големи стойности на явлението “фактор” () отговарят по-големи стойности на явлението “**резултат**” ().



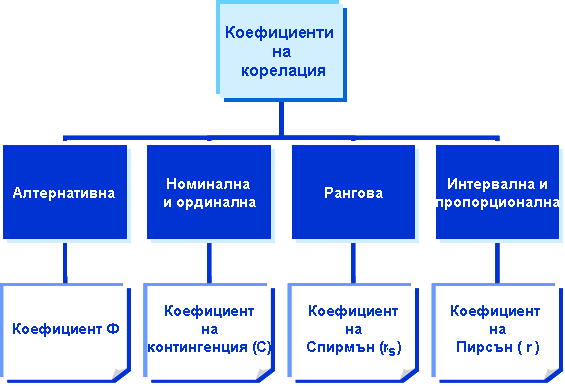
При зависимостта е низходяща (разнопосочна) – нарастването на стойностите на води до намаляване на стойностите на .

Логиката на корелационните коефициенти може да бъде пояснена и по друг начин. Стойностите на зависимата променлива () варират под влияние на различни фактори. Една част от тяхното разсейване се обяснява с влиянието на независимата променлива (), а останалата – с влиянието на други фактори. Колкото по-силна е зависимостта между променливите – толкова по-голяма част от дисперсията на стойностите на се дължи на влиянието на , а съответно по-малка – на неизследваните фактори.

**Коефициентът на определеност (детерминираност) (**), който е равен на квадрата на коефициента на корелация, описва т. нар. „обяснена дисперсия“. Той показва каква част от вариацията на се дължи на различията в стойностите на , т.е. на влиянието на изучавания фактор. Ако се умножи по 100, изразява силата на влияние на зависимата променлива в проценти.

**Коефициентът на неопределеност ()** описва влиянието на невключените в изследването фактори, на които се дължи т. нар. необяснена дисперсия. Стойността му се получава, като от общата дисперсия на променливата Y, изразена като единица (100%), се извади обяснената с помощта на коефициента на детерминираност **(**). С други думи Очевидно е, че колкото по-силна е зависимостта между променливите величини, толкова обяснената дисперсия е по-голяма, а необяснената – по-малка.

В зависимост от вида и формата на зависимостта, както и от начина на скалиране на променливите величини, в практиката се използват различни коефициенти на корелация. На **фиг. 4.2** е направен преглед на най-често използваните коефициенти.



**Фиг. 4.2.**

### Коефициент на обикновена линейна корелация на Пирсън (r)

Прилага се прилага в случай, че зависимостта е:

* обикновена по вид, т.е. изследва се връзката между две променливи величини.
* линейна по форма, т. е. предварителната проверка е доказала линейния характер на връзката между променливите.
* признаците и са представени в пропорционална или интервална скала.

Стойността на коефициента на Пирсън (r) се изчислява по следната формула:

* + е момент на произведенията;
  + е стандартното отклонение на променливата ;
  + е стандартното отклонение на променливата ;

Моментът на произведенията (Р) се изчислява по формула:

* + е сумата на стойностите на ;
  + е сумата на стойностите на ;
  + е сумата на произведенията на ;

### Коефициент на рангова корелация на Спирман ()

Използването на този коефициент е коректно, когато:

* двете променливи са рангово скалирани признаци.
* едната променлива е рангова, а другата количествена. В такъв случай предварително е необходимо количествената променлива да бъде трансформирана в рангова скала.
* количествени променливи имат не-нормално разпределение.

Стойността на коефициента се изчислява по формулата:

* + обем на извадката;
  + е разликата в ранговите номера по ;

## Статистическа значимост на коефициентите на корелация

При статистически проучвания изследователят се интересува от проявлението на зависимостите между явленията. Информацията, с която разполага, обикновено се базира на данни от извадка, а това означава, че стойността на коефициента на корелация може да бъде повлияна от случайни фактори.

Поради това се налага проверка на хипотезата за статистическата значимост на корелационния коефициент. Нулевата хипотеза гласи, че зависимост между изучаваните явления обективно не съществува, т.е. , следователно емпиричният коефициент може да се счита за случаен.

Ето как трябва да постъпим, ако се използват таблици с критични стойности на следните коефициенти:

* Коефициент на обикновена линейна корелация на Пирсън () при степени на свобода и равнище на значимост и 0,01.
* Коефициент на рангова корелация на Спирман () при степени на свобода и равнище на значимост и 0,01.

В случай, че стойността на изчисления по данни от извадката коефициент на корелация е по-малък от критичната стойност, за вярна се приема нулевата хипотеза, която твърди, че няма зависимост между променливите величини.

## Крос-корелация (примери с 2 сигнала)

Корелация с изместване по времето…

<https://nl.mathworks.com/help/signal/ug/align-signals-using-cross-correlation.html>

<https://en.wikipedia.org/wiki/Cross-correlation>

# Параметрични и непараметрични критерии

## Фактори, изследвани с параметрични и непараметрични критерии.

На практика, за да разберем идеята на непараметричната статистика, първо трябва да обясним базовите концепции на параметричните статистики.

Идеята за тестове за статистическа значимост е базирана на примерни (извадкови) разпределения на дадена статистика. Накратко, ако имаме базови знания за разпределението на дадена случайна величина (променлива, фактор), то можем да направим предположение за това как (в повтарящи се извадки с един и същи размер) ще се държи нашата статистика, т.е. как е разпределена наблюдаваната величина. Например, ако направим 100 случайни извадки от по 100 възрастни (включваме цялото население на земята) и изчислим средната височина на индивидите във всяка извадка, то разпределението на стандартното средно в извадките ще бъде много близко до нормалното разпределение (да бъдем по-точни разпределението на Стюдънт – разпределение с 99 степени на свобода). Сега да си представим, че направим една допълнителна извадка в точно определен град София, където предполагаме, че хората са по-високи от средностатистическия жител на земята. Ако средната височина на хората от извадката е извън зоната на горните 95% на t – разпределението, то можем да заключим, че наистина хората в София са по-високи от средното за населението.

## Дали повечето случайни променливи (величини) са нормално разпределени?

В горния пример се опряхме на нашето знание, че при последователни извадки с еднакъв размер, стандартното средно (в нашия пример за височината) ще бъде разпределено следвайки t – разпределението (с конкретно средно и вариация). Все пак това би било единствено истина, ако за населението, променливата, от която се интересуваме (в нашия случай височината) е нормално разпределена. За повечето променливи, от които се интересуваме, просто не знаем със сигурност дали това е така. Например появяването на редки заболявания сред населението или пък броят на пътните инциденти и много други интересни за изследване величини или не са нормално разпределени или нямаме достатъчно данни за да го установим.

## Размерът на извадката

Друг фактор, който често лимитира приложимостта на тестовете, базирани на предположението (допускането) че извадковото разпределение е нормално е размерът на извадката от данни, достъпни за анализ. Можем да допуснем, че разпределението на извадката е нормално само за достатъчно големи извадки (100 и повече наблюдения). Ако обаче нашата извадка е малка, можем да използваме тестовете само ако сме сигурни, че тя е нормално разпределена и нямаме никакъв начин да тестваме нашето предположение.

## Проблемът с измерванията

Прилагането на тестове, базирани на предположението за нормалност на разпределението е допълнително ограничено от липсата на прецизно измерване. Например, нека разгледаме изследване, в което средната оценка от изпити (СОИ)е приета за основна променлива, която ни интересува. Дали средното А е два пъти по-добро от средното Б? Дали разликата между средните А и Б е сравнима с разликата между средните В и Г? По някакъв начин (СОИ) е сурова мярка за завършване на училище, която само ни позволява да подредим учениците от слаб до отличен. Без да навлизаме в дълбочина, най-общите статистически техники като анализ на вариациите (и t-тестове), регресия и т.н. предполагат, че измерванията са поне от интервално качество, което означава, че можем да сравняваме резултатите от функции, действащи върху измерваните величини.

## Параметрични и непараметрични методи

Считаме, че след това въведение е очевидна нуждата от статистически методи, позволяващи да обработваме данни с „лошо качество”, от малки извадки, за променливи, за които нищо не се знае предварително (по отношение на тяхното разпределение). Непараметричните методи са специално разработени да се използват в случаи, в които изследователят не знае нищо за параметрите на променливата, която го интересува (от тук и името – непараметрични). Научно погледнато, непараметричните методи и критерии не разчитат на оценка на параметрите (такива, като средно или стандартно отклонение) описващи разпределението на интересуващата ни променлива. Всеки параметричен тест има поне един непараметричен еквивалент. Основно тези тестове попадат в някоя от следните категории:

* Тестване на разликите между групи (независими извадки);
* Тестване на разликите между групи (зависими извадки);
* Тестване на връзки между променливите;

# Проверка на хипотези при независими извадки

## Хипотеза и алтернатива

Голяма част от научните изследвания се отнасят до сравняване на разпределението на две или повече променливи величини.

Особеното на тези сравнения е, че изводите, които се правят, трябва да се отнасят до целите съвкупности, а данните, с които разполага изследователят, обхващат само извадка от нея. Поради това първоначално се формулират предположения – хипотези, и се прави проверка дали данните от извадката ги потвърждават или отхвърлят. Така например отговорът на въпроса има ли разлика в ефекта на две методики може да бъде сведен до две предположения – “не, няма разлика в ефекта им” и “да, има разлика”. Тези две възможности изразяват съдържанието на двете статистически хипотези:

* **Нулева или работна (Н0)** - тя твърди, че няма статистически достоверна разлика в сравняваните статистически показатели. Въпреки, че в извадките може да се наблюдава известна разлика, тя е случайна и не може да бъде обобщена за генералните съвкупности.
* **Алтернативна хипотеза (Н1)**, която твърди, че констатираната разлика в емпиричните данни е статистически достоверна и може да бъде обобщена за генералните съвкупности.

Решенията, които се вземат при проверката на статистически хипотези, имат вероятностен характер. Това се дължи на факта, че **изследванията са репрезентативни**, т.е. изводите за съвкупността се правят въз основа на изследване на относително малка част от нея. Приемането или отхвърлянето на нулевата хипотеза се прави с определена сигурност, като в същото време се допуска възможността за грешка.

Степента на сигурност, с която се приема за вярна алтернативната хипотеза, се нарича г**аранционна вероятност ().**

Рискът да се допусне грешка, като се приеме за вярна алтернативната хипотеза се нарича **равнище на значимост ()** в математиката и *p-value* в статистиката.

В практиката се използват следните стойности за гаранционната вероятност () и равнище на значимост (****):

* , на която съответства (5% възможност за грешка).
* , на която съответства (1% възможност за грешка).
* , на която съответства (0,1% възможност за грешка).

В случай, че , съответно , за вярна се приема нулевата хипотеза. В противен случай – алтернативната.

## Поредица от стъпки за проверка на хипотези

Общата процедура за проверка на статистически хипотези преминава през няколко поредни стъпки:

***І стъпка*** *-* ***Формулиране на нулевата и алтернативната хипотеза.***

Както вече бе споменато, нулевата (H0) хипотеза гласи, че няма разлика в сравняваните статистически показатели. Ако емпиричните данни сочат известни различия, то те се дължат на случайни фактори. Противоположно е твърдението на алтернативната хипотеза, според която наблюдаваните в емпиричните данни различия (ефект, зависимост) са резултат на закономерно действащи фактори.

***ІІ стъпка - Избор на подходящ критерий за проверка на хипотезата.***

Хипотезите се проверяват със специфични критерии. За целта е необходимо да се познават условията, на които трябва да отговарят променливите величини, броят и видът на извадките и някои други особености, които са разгледани в следващия раздел.

***ІІІ стъпка - Изчисляване на емпиричната стойност на критерия (Критерий emp).***

За всеки един критерий има разработена процедура за изчисляването му по данни от извадката.

***ІV стъпка – Определяне на табличната стойност на критерия (Критерий ).***

Тя се взема от приложение, в зависимост от степените на свобода **(k)** и равнището на значимост **()**. Към статистическата литература има приложени таблици с критични стойности на критериите.

***V стъпка – Вземане на решение.***

За целта табличната (критичната) стойност на критерия се сравнява с емпиричната (изчислената по данни от извадката). Именно това сравнение дава основание да се направи извод коя от хипотезите (нулевата или алтернативната) е вярна. Възможностите са две:

* Ако емпиричната стойност е по-малка от табличната за вярна се приема нулевата хипотеза (H0).
* Ако емпиричната стойност е по-голяма или равна на табличната - нулевата хипотеза (H0) се отхвърля и за вярна се приема алтернативната хипотеза (Н1).

Вземането на решение може да стане и въз основа на равнището на значимост **()**, което съответства на емпиричната стойност на критерия, чиято стойност се изписва в разпечатките например на SPSS:

* **Ако равнището на значимост ()** е по-голямо от възприетото ( > 0,05 (0,01 или 0,001) - няма основание за отхвърляне на нулевата хипотеза (Н**о**)
* **Ако равнището на значимост ()** е по-малко или равно на възприетото ( ≤ 0,05 (0,01 или 0,001) - нулевата хипотеза (Н**о**) се отхвърля в полза на алтернативната.

Както вече бе споменато, аналогична информация носи гаранционната вероятност P/%/, която при известно равнище на значимост се изчислява по формулата:

В такъв случай:

* Ако гаранционната вероятност е по-малка от възприетата (95% или 99%) - за вярна се приема нулевата хипотеза (**H0**).
* Ако гаранционната вероятност е равна или по-голяма от 95% (99%) - нулевата хипотеза се отхвърля и за вярна се приема алтернативната хипотеза.

## Критерии за проверка на хипотези

Изборът на критерий за проверка на дадена хипотезата зависи от:

**1. Използвания статистически метод** и показатели, които се сравняват (за сравняване на средни аритметични, дисперсии, относителни дялове и др. се ползват различни критерии).

**2. Естеството на променливите величини**, които подлежат на обработка – на първо място е важно дали променливите са количествени или категорийни (качествени). Вторият фактор е нормално ли е разпределението на признака или не. От тази гледна точка са разработени две големи групи от статистически критерии:

* **Параметрични,** които се прилагат при количествени (интервално и пропорционално скалирани) признаци, когато предварително е доказано, че разпределението е нормално.
* **Непараметрични** – прилагат се при качествени (номинално и рангово) скалирани признаци, а така също, когато разпределението на количествен признак не е известно или не е нормално.

**3. Вида на извадките, които подлежат на сравнение.** От гледна точка на начина, по който са съставени, те биват:

* **Независими** - когато изследваните лица от двете извадки са различни (сравняване на експериментална и контролна група, мъже-жени и др.)
* **Зависими** са извадките, когато единиците от едната извадка предопределят тези във втората. Типичен случай са изследванията за сравняване на променливи преди и след някакво въздействие. Очевидно е, че експерименти от такова естество са коректни само в случай, че изследваните лица (единици) от първото и второто изследване са едни и същи.

**4. Брой извадки, които се сравняват.**

Очевиден критерий

*В таблица 5.1 е направена систематизация на основите критерии*

таблица 5.1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Група** | **Променливи величини** | **Брой на извадките** | **Вид на извадките** | **Критерии за проверка на хипотези** |
| **Параметрични** | **Количествени, които имат нормално разпределение** | Две | Независими | -критерий на Стюдънт за независими извадки |
| Зависими | -критерий на Стюдънт за зависими извадки |
| Три и повече | Независими  Зависими | -критерий на Фишер (дисперсионен анализ)  -критерий на Фишер (дисперсионен анализ за повтарящи се наблюдения) |
| **Непараметрични** | **Алтернативни** | Две | Независими | -критерий на Стюдънт |
| Зависими | -критерий на Мак Немар |
| **Номинални** | Две и повече | - | -критерий на Пирсън |
| **Рангови и количествени, които нямат нормално разпределение** | Две | Независими | -критерий на Ман Уитни |
| Зависими | -критерий на Уилкоксън |
| Три и повече | Независими | Критерий на Крускал Уолис |
| Зависими | **-** критерий наФридман |

## Сравняване на средно равнище на признака при две независими извадки. t -критерий на Стюдънт за независими извадки

Той се прилага за проверка на нулевата хипотеза за сравняване на средни аритметични величини , когато:

* Признаците са количествено измерими.
* Имат нормално разпределение.
* Извадките са две и са независими.

**Нулевата хипотеза** твърди, че няма статистически достоверна разлика в сравняваните средни стойности.

**Емпиричната стойност** на критерия се изчислява (temp*)* по формула:

* са средните на извадките;
* са дисперсии в извадките;
* обем на извадките;

Определя се **таблична стойност на критерия** (**t**) , в зависимост от равнището на значимост (****) и степените на свобода **k=*n1*+ *п2*– 2**.

Проверява се верността на нулевата хипотеза по условието:

* Ако **temp**< **t** (** >0,05** или **P/t/<95%)** е вярна нулевата хипотеза.
* При **temp** **≥** **t** (**≤ 0,05** или **P/t/** **≥95%)** е вярна алтернативната

## Сравняване на средно равнище на признака на три и повече независими извадки. F-критерий на Фишер (еднофакторен дисперсионен анализ)

Дисперсионният анализ представлява метод за проверка на хипотезата за сравняване на средното равнище на признака на три и повече извадки. Резултатът от сравнението дава възможност да се прави заключение дали даден фактор влияе върху изследвания признак.

Прилагането на дисперсионния анализ е коректно когато:

* Променливата, която се изследва, е количествена, а факторът - изразен в номинална или ординална скала е пропорционално скалирана променлива, а другата - номинално скалирана.
* Извадките са независими.
* Разпределението на признака е нормално и с приблизително еднакви дисперсии.

Логиката на дисперсионния анализ се основава на разлагане на общата вариация (дисперсия) на променливата (отклоненията на всички стойности от тяхната обща средна) на две важни компоненти:

* **Вътрешногрупова дисперсия,** изразяваща се във индивидуални отклонения на стойностите от средната в рамките на дадена категория, група.
* **Междугрупова дисперсия** - отклоненията на груповите средни от общата средна аритметична.

Колкото по-силно действа факторният признак, толкова по-големи ще са разликите между средните аритметични величини на групите и междугруповата вариация.

**Емпиричната стойност** на F–критерия на Фишер се изчислява по формула:

* + е между-групова дисперсия;
  + е вътрешно-групова дисперсия;

**Таблична стойност** на критерия **F** се отчита в зависимост от равнището на значимост **** =0,05 и степени на свобода както следва:

* за между-груповата дисперсия k1=m–1 (където m е броят на сравняваните групи).
* за вътрешно-груповата дисперсия k2=n - m.

**Проверява се верността на нулевата хипотеза по условието:**

* Ако **Femp**< **F** (** >0,05** или **P/F/<95%)** е вярна нулевата хипотеза.
* При **Femp** **≥** **F** (**≤ 0,05** или **P/F/** **≥95%)** е вярна алтернативната

Установяването на верността на алтернативната хипотеза не уточнява между кои средни се наблюдава достоверна разлика. Това означава, че средните на групите трябва да се сравнят по двойки с t-критерия на Стюдънт за независими извадки.

# Зависими извадки

## Показатели за описание на развитие

Най-често, обработката на резултати от педагогически експеримент предполага да се проследи прираста на резултатите между две последователни изследвания (преди и след). Показателите, с които описват настъпилите промени са:

абсолютен прираст

относителен прираст

Статистическата достоверност на прираста се проверява с t-критерия на Стюдънт за зависими извадки.

## Сравняване на средни величини при зависими извадки (t-критерий на Стюдънт за зависими извадки)

Този статистически критерий се използва за сравняване на средни аритметични величини, когато:

* Признаците са метрично скалирани и имат нормално разпределение.
* Извадките са две и са зависими.

Проверката на нулевата хипотеза става в следната последователност:

* **Емпиричната стойност** на t-критерия на Стюдънт за зависими извадки се изчислява по формула:
  + е разлики между І и ІІ изследване за всяко изследвано лице;
* Определя се **таблична стойност на критерия** (**t**) от приложение 6, в зависимост от равнището на значимост (****) и степените на свобода **k=*n*– 1**.
* Проверява се верността на нулевата хипотеза по условието:
  + Ако **temp**< **t** (** >0,05** или **P/t/<95%)** е вярна нулевата хипотеза
  + При **temp** **≥** **t** (**≤ 0,05** или **P/t/** **≥95%)** е вярна алтернативната хипотеза

## Приложение на проверката на хипотези при обработка на данни от педагогически експеримент

Един от най-разпространените методи за научни изследвания е педагогическият експеримент. Целта на провеждането му е да се проследи ефектът на някакво ново въздействие. В такива случаи обикновено възникват два изследователски проблема:

* Има ли ефект новото въздействие?
* По-висока ли е ефективността му в сравнение с традиционните, внедрени вече в практиката въздействия?

За решаване на тези две изследователски задачи обикновено се формират две групи:

* **Експериментална** - с която се прилага новото въздействие;
* **Контролна** - с която се прилага традиционното въздействие.

За проследяване на ефекта от тези въздействия се провеждат най-малко две изследвания:

* **Начално** - за отчитане на равнището на признака преди въздействието.
* **Крайно** - за отчитане на равнището на признака след въздействието.

Интерпретацията на резултатите, получени от експеримента, протича в следната логическата последователност:

**1.** **На първо място се установява има ли ефект новото въздействие.** За целта се сравнява средното равнище на признака при експерименталната група в началото и в края на експеримента. Описва се величината на настъпилите промени (с **d** и **d%**) и се установява неговата статистическа достоверност. Тъй като се сравняват средни стойности на променливи, измерени при едни и същи лица се ползва t-критерият на Стюдънт за зависими извадки. По аналогичен начин се сравняват средните в началото и в края на експеримента при контролната група.

**2. Следващият важен момент е съпоставяне на ефекта, постигнат в експерименталната и контролната група.** За да бъде коректно това сравнение е необходимо предварително да се докаже, че при започване на експеримента двете групи са били на едно и също ниво. За целта е необходимо да се сравни средното равнище на признака на експерименталната и контролната група в началото на експеримента и се установи има ли статистически значима разлика между тях. Тъй като лицата в двете групи са различни, сравнението се осъществява с t-критерия на Стюдънт за независими извадки.

**3.** В случай, че това условие е изпълнено, се преминава към следните сравнения:

* На **крайните резултати** на експерименталната и контролната група.
* На **прирастите в двете групи** (**de** и **dk**). За целта е необходимо да се образуват производни променливи като разлика между начално и крайно състояние на признака и да се сравнят средните аритметични на прираста на групите. Статистическата достоверност на различията на този етап от интерпретацията на резултатите от педагогическия експеримент се установява с -критерия на Стюдънт за независими извадки.

**References**

1. *"Fairness Opinions: Common Errors and Omissions". The Handbook of Business Valuation and Intellectual Property Analysis. McGraw Hill. 2004.*[*ISBN*](https://en.wikipedia.org/wiki/ISBN_(identifier))[*0-07-142967-0*](https://en.wikipedia.org/wiki/Special:BookSources/0-07-142967-0)*.*
2. *Agrrawal, Pankaj; Borgman, Richard; Clark, John M.; Strong, Robert (2010). "Using the Price-to-Earnings Harmonic Mean to Improve Firm Valuation Estimates".*[*Journal of Financial Education*](https://en.wikipedia.org/w/index.php?title=Journal_of_Financial_Education&action=edit&redlink=1)*.****36****(3–4): 98–110.*