

СЪДЪРЖАНИЕ.

ТЕОРИЯ НА ВЕРОЯТНОСТИТЕ

Вероятност на случайно събитие.

- §1. Случаен експеримент и случайно събитие. Действия със събития. (3)
- §2. Относителна честота и вероятност на случайно събитие. Условна вероятност. (6)
- §3. Статистическа, класическа и геометрична вероятност. (10)
- §4. Елементи от комбинаториката. (13)
- §5. Формула за пълната вероятност. Формули на Бейс. (17)

Случайни величини. (23)

- §6. Случайна величина. Видове. Функция на разпределение на случайна величина. (23)
- §7. Дискретни случайни величини. (25)
- §8. Математическо очакване и други числени характеристики на дискретна случайна величина. (29)
- §9. Непрекъсната случайна величина. Функция и плътност на разпределение. Мода, медиана, квантили и критични точки. (33)
- §10. Математическо очакване, дисперсия и други числени характеристики на непрекъснати случайни величини. (38)

Някои по-важни дискретни и непрекъснати разпределения. (44)

- §11. Биномно разпределение, разпределение на Пуасон и геометрично разпределение. (43)
- §12. Равномерно и показателно разпределение. Сума на две и повече случайни величини. (49)
- §13. Нормален закон на разпределение. (52)
- §14. Други разпределения на непрекъснати случайни величини. (55)
- §15. Гранични теореми. (63)
- §16. Неравенство на Чебишев. Закон за големите числа. (67)

Двумерни случайни величини. (71)

- §17. Двумерна случайна величина - закон за разпределение. Закони за разпределение на съставните величини. (71)
- §18. Условни разпределения. Условно математическо очакване. Регресия и линия на регресията. (76)
- §19. Ковариация. Коефициент на корелация. (80)

СТАТИСТИКА

- §20. Предмет и задачи на математическата статистиката. Генерална съвкупност и извадка. (83)
- Статистическа обработка на данни. (85)**
- §21. Статистическо разпределение, полигон и хистограма. Емпирична функция на разпределение. (85)
- §22. Числени характеристики на статистическото разпределение. Характеристики на местоположението - мода, медиана, квантили и квартили. (92)
- §23. Средна, дисперсия и средно квадратично отклонение, начални и централни моменти на извадката. (98)

Оценка на параметри. (103)

- §24. Точкови оценки (статистики). Видове оценки – неизместена и изместена, състоятелна и ефективна. (103)
- §25. Точкови оценки на генералното математическо очакване и дисперсия. (106)
- §26. Начини за намиране на оценки на параметри - метод на моментите. Метод на най-голямото правдоподобие. (108)
- §27. Интервални оценки. Доверителен интервал и доверителна вероятност. (111)
- §28. Доверителни интервали за математическото очакване и дисперсията на нормално разпределена случайна величина. (113)
- Проверка на хипотези. (118)**
- §29. Статистически хипотези. Нулева и конкурираща хипотеза. Грешка от първи и грешка от втори род. Статистически критерий. (118)
- §30. Непараметрични тестове - χ^2 - критерий на Пирсън за вида на разпределението. (124)
- §31. Параметрични тестове - тестове за математическото очакване и дисперсията на нормално разпределена генерална съвкупност. (128)
- Изследване на две извадки. (137)**
- §32. Зависими и независими извадки. Изследване на две зависими извадки. (137)
- §33. Доверителни интервали и критерии за сравняване на дисперсиите и математическите очаквания на две нормално разпределени генерални съвкупности. (139)
- Корелационен и регресионен анализ. (146)**
- §34. Двумерни извадки. Числени характеристики на извадката и точкови оценки. Корелационна таблица. (146)
- §35. Елементи на корелационния анализ. Видове зависимости между две величини. (151)
- §36. Елементи на регресионния анализ. Метод на най-малките квадрати за намиране на линията на регресия. (156)
- §37. Статистически анализ на уравнението на линейната регресия. (161)
- Приложения. (168)**

ВЕРОЯТНОСТ НА СЛУЧАЙНО СЪБИТИЕ

§1. Случен експеримент и случайно събитие. Действия със събития.

Теорията на вероятностите е математическа дисциплина, изучаваща закономерностите на случайните явления. Едно от основните понятия в нея е случайното събитие – всяко явление, което при едни и същи условия може да се осъществи, а може и да не се осъществи. Със случайни събития се срещаме навсякъде – такива са, например, повреда на техническо устройство, шум при приемане на съобщение, печалба от лотария и др.

Един експеримент (опит) се състои в осъществяване на съвкупност от точно определени условия и отчитане на произтичащите от тях събития, които накратко ще наричаме изходи. Всяко осъществяване на съвкупността от условията наричаме реализация на опита (експеримента). В теорията на вероятностите, се разглеждат само такива експерименти, които могат да се повтарят произволен брой пъти, поне теоретически.

Един експеримент се нарича случаен (стохастичен, вероятностен), ако при непроменени условия е възможно настъпването на взаимно изключващи се изходи. Всеки наблюдаван резултат от случаен експеримент ще наричаме случайно събитие.

Пример 1.1. В една кутия има 4 еднакви топки, които са номерирани 1, 2, 3 и 4. Опитът се състои в изтегляне по случаен начин на две топки без връщане. Възможни са 6 случая (изходи), които ще означаваме накратко

$$\omega_1 = \{1,2\}, \omega_2 = \{1,3\}, \omega_3 = \{1,4\}, \omega_4 = \{2,3\}, \omega_5 = \{2,4\}, \omega_6 = \{3,4\} \quad (1.1)$$

като в резултат винаги настъпва точно един от тях. Например, събитието $A = \{\text{изтеглени топки с номера, по-малки от 3}\}$ се сбъдва само, ако резултатът от опита е ω_1 , и не настъпва, ако имаме някой от изходите $\omega_2, \omega_3, \omega_4, \omega_5$ или ω_6 ; събитието $B = \{\text{изтеглени са две топки}\}$ винаги настъпва, докато събитието $C = \{\text{изтеглени са три топки}\}$ в този опит е невъзможно. ♦

Събитие, което се сбъдва при всяка реализация на опита, се нарича достоверно. Събитие, което не настъпва при нито една реализация на опита, се нарича невъзможно. Ако събитието A не е настъпило, казваме, че е настъпило противоположното му събитие $\bar{A}$.

Противоположното на достоверното събитие е невъзможното.

Казваме, че събитието A влече събитието B (B е следствие от A), което означаваме с $A \subset B$, ако всеки път, когато настъпва събитието A , настъпва и събитието B .

Например, ако е настъпило събитието A от пример 1.1, то е настъпило и събитието $D = \{\text{сумата от номерата на изтеглените топки е}$

по-малка от 5}, т.е. $A \subset D$. Очевидно е, че за този пример са изпълнени също и отношенията $A \subset B, D \subset B$.

Ако $A \subset B$ и $B \subset A$, то казваме, че събитията A и B са еквивалентни и пишем $A = B$. В пример 1.1 $A = \omega_1$.

Казваме, че две събития A и B са несъвместими, ако в никоя реализация на опита не е възможно да появата и на двете събития. Очевидно, всяко събитие и неговото противоположно са несъвместими събития. В пример 1.1 събитията C и B също са несъвместими.

Казваме, че събитията $A_1, A_2, \dots, A_n$ образуват пълна група несъвместими събития, ако две по две са несъвместими и в резултат на проведен опит винаги настъпва някое от тях.

Всяко събитие и неговото противоположно образуват пълна група несъвместими събития.

Изходите $\omega_1, \dots, \omega_6$ на опита в пример 1.1 са пример за пълна група несъвместими събития, при това не е трудно да се убедим, че всяко събитие, което е свързано с този опит, може да се изрази чрез тях. Например, $A = \{\omega_1\}$, $\bar{A} = \{\omega_2, \dots, \omega_6\}$, $D = \{\omega_1, \omega_2, \omega_3\}$, т.е. $A, \bar{A}$ и D са подмножества на множеството $\Omega = \{\omega_1, \dots, \omega_6\}$, състоящо се от всички възможни изхода на експеримента. По подобен начин всяко случайно събитие, наблюдавано в този опит, може да се разглежда като подмножество на множеството Ω .

Всеки изход ω на един експеримент наричаме елементарно събитие. Множеството Ω от всички възможни елементарни събития наричаме пространство на елементарните събития на експеримента.

Всяко случайно събитие A се явява подмножество на Ω , т.е. $A \subseteq \Omega$, елементите на което са елементарните събития, при които настъпва случайното събитие. Тези елементарни събития се наричат благоприятни за A .

Достоверното събитие ще означаваме с Ω , тъй като благоприятно за него е всяко елементарно събитие, а невъзможното ще означаваме с $\emptyset$.

Действия със събития. Нека са дадени две събития A и B . От тях образуваме нови събития, които ще наричаме:

- Сума (обединение) $A \cup B$ ($A+B$, A или B), ако настъпи поне едно от събитията A и B .
- Произведение (сечение) $A \cap B$ (AB , A и B), събитие, което се състои в съвместно настъпване на A и B .
- Разлика $B \setminus A$ - събитие, състоящо се в сбъждане на събитието $B \cap \bar{A}$, т.е. настъпва събитието B , но не настъпва събитието A .

Пример 1.2. Опитът се състои в подхвърляне на една монета два пъти. Разглеждаме събитията $A=\{\text{при първото подхвърляне се пада лице}\}$, $B=\{\text{при второто подхвърляне се пада лице}\}$. а) Да се изрази чрез събитията A , B , $\bar{A}$ и $\bar{B}$ събитието $C=\{\text{при двата опита се падат различни страни на монетата}\}$ и противоположното му събитие. б) Да се състави пространството Ω на елементарните събития за този опит и да се изразят чрез тях събитията A , B , C , $A \cup B$, $A \cap B$ и техните противоположни.

Решение.

а) Събитието C настъпва, когато или първия път се пада лице, (събитието A), а втория път не се пада лице (събитието $\bar{B}$) или обратното – първия път се пада герб $\bar{A}$, а втория – лице (B), т.е. $C=(A \cap \bar{B}) \cup (\bar{A} \cap B)$. Събитието $\bar{C}$ настъпва, ако двата пъти монетата се е обърнала от една и съща страна, което се изразява като $\bar{C}=(A \cap B) \cup (\bar{A} \cap \bar{B})$.

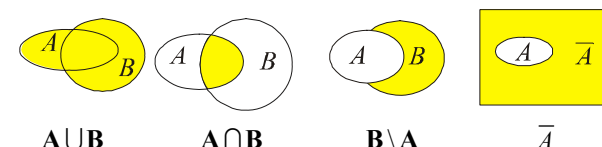
б) $\Omega=\{ЛЛ, ЛГ, ГЛ, ГГ\}$. Събитието A настъпва, ако са настъпили изходите $ЛЛ$ и $ЛГ$, т.е. A се представя чрез подмножеството $A=\{ЛЛ, ЛГ\}$ на Ω . Аналогично, подмножествата с които се представят останалите събития са (изкажете с думи събитията):

$\bar{A}=\{ГЛ, ГГ\}$, $B=\{ГЛ, ЛЛ\}$, $\bar{B}=\{ГГ, ЛГ\}$, $C=\{ЛГ, ГЛ\}$, $\bar{C}=\{ГГ, ЛЛ\}$
 $A \cup B=\{ЛЛ, ЛГ, ГЛ\}$, (всички благоприятни случаи за събитията A и B)
 $A \cap B=\{ЛЛ\}$, (единственият общ случай, при който настъпват A и B)
 $A \setminus B=\{ЛГ\}$, $B \setminus A=\{ГЛ\}$. ♦

Както се вижда от примера, ако Ω е пространство от елементарни събития за даден опит и събитията A и B се разглеждат като подмножества A и B на Ω , т.е. $B \subset \Omega$, $A \subset \Omega$, то

- сумата на събитията A и B съвпада с обединението $A \cup B$, което се състои от елементите, принадлежащи поне на едно от множествата A и B .
- произведението на събитията A и B съвпада с сечението $A \cap B$, което е съставено от всички елементи, принадлежащи както на A , така и на B .
- разлика $B \setminus A$ на събитията A и B съвпада с разликата $B \setminus A$ на множествата A и B , т.е. множеството състоящо се от всички елементи на B , които не принадлежат на A .

Действията с множества са представени по-долу чрез множества от точки в равнината:



Ще цитираме някои основни равенства, отразяващи основните свойства на въведените действия, които са в сила както за множества, така и за събития:

$$1) A \cup \emptyset = \emptyset \cup A = A, \quad 2) A \cap \emptyset = \emptyset \cap A = \emptyset$$

$$3) A \cup \bar{A} = \Omega, \quad 4) A \cap \bar{A} = \emptyset,$$

Комутативни закони:

$$5) A \cup B = B \cup A; \quad 6) A \cap B = B \cap A$$

Асоциативни закони (независимост от реда на извършване на действията)

$$7) A \cup (B \cap C) = (A \cup B) \cap C = A \cup B \cap C,$$

$$8) A \cap (B \cup C) = (A \cap B) \cup C = A \cap B \cup C$$

Дистрибутивни закони (разкриване на скоби):

$$9) A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \quad 10) A \cup (B \cap C) = (A \cup B) \cap (A \cup C),$$

Формули на Де Морган:

$$11) \overline{A \cup B} = \bar{A} \cap \bar{B}, \quad 12) \overline{A \cap B} = \bar{A} \cup \bar{B}.$$

Упражнения.

1. Монета се подхвърля до първата поява на герба. Да се опише пространството от елементарните събития. Да се изразят събитията $A=\{\text{герб се пада при третото подхвърляне}\}$, $B=\{\text{герб се пада не по-рано от третото подхвърляне}\}$.
2. Опитът се състои в подхвърляне три пъти на една монета. Пространството от елементарни събития е $\Omega=\{ЛЛЛ, ЛЛГ, ЛГЛ, ГЛЛ, ГЛГ, ГЛГ, ГЛГ, ГГГ\}$. а) Да се опишат с думи събитията $A=\{ЛЛЛ, ЛЛГ, ЛГЛ, ЛГГ\}$, $B=\{ЛЛЛ, ГГГ\}$, $C=\{ЛЛГ, ЛГЛ, ГЛЛ\}$, $D=\{ЛЛГ, ЛГЛ, ГЛЛ, ГГЛ, ГЛГ, ЛГГ, ГГГ\}$ и техните противоположни събития. б) Кои от събитията са следствие на събитието C ?

§2. Относителна честота и вероятност на случайно събитие. Условна вероятност.

Известно е, че някои случайни събития се случват по-често, а други – по-рядко. Например, много по-рядко се печели шестца от тотото, отколкото четворка. Затова възниква необходимостта да се въведе някаква мярка, която ще наричаме *вероятност* – реално число, което да характеризира колко често настъпва дадено събитие. По неговата стойност можем да преценим възможността за настъпване на дадено случайно събитие, да сравним шансовете за появата различни събития, т.е. да познаваме по-цялостно случайните експерименти и явления.

Естествен път за определяне колко често настъпва едно събитие A в даден случаен експеримент е да се направи серия от опити и да се преброи колко пъти е настъпило събитието.

Относителна честота на събитието A наричаме отношението

$w(A) = \frac{m}{n}$, където m е броят на поява на A в серия от n опита.

Доказва се, че относителната честота има свойствата:

- $0 \leq w \leq 1$, $w(\Omega) = 1$, $w(\emptyset) = 0$,
- $w(A \cup B) = w(A) + w(B)$, ако A и B са несъвместими събития,
- $w(B|A) = \frac{w(A \cap B)}{w(A)}$, където с $w(B|A)$ е означена относителната

честота на събитието B , при условие, че събитието A е настъпило.

Пример 2.1. Естествоизпитателят Буфон, подхвърлял една монета 4040 пъти като резултатът бил 2048 пъти се пада лице (събитие A) и 1992 пъти герб (събитие B), т.е. относителните честоти на тези събития са $w(A) = \frac{2048}{4040} \approx 0,5069$, $w(B) = \frac{1992}{4040} \approx 0,4931$, откъдето може да се заключи, че двете събития имат почти равни шансове за поява. ♦

Очевидно, относителната честота дава представа колко често настъпва дадено събитие, но все пак зависи от серията, от която е изчислена. Затова статистически определената честота на събитието не носи строг характер, а дава само практически указания какви свойства би трябвало да има вероятността на едно събитие, че да служи като количествена мярка на обективната възможност за сбъждане на събитието. Изходни свойства в определението за вероятност се вземат свойства 1) и 2) на относителната честота. Тук те не се доказват, а се приемат като задължителни, затова са наречени **аксиоми**. (Строгата дефиниция за вероятност на събития, която за пръв път е дадена от Андрей Колмогоров, няма да разглеждаме).

На всяко случайно събитие A , наблюдавано при даден случаен експеримент, се съпоставя число $P(A)$, наречено вероятност на събитието A , със свойствата:

Аксиома 1. $0 \leq P(A) \leq 1$.

Аксиома 2. $P(\Omega) = 1$.

Аксиома 3. За произволна крайна или безкрайна редица от наблюдавани събития $A_1, A_2, \dots, A_n, \dots$, такива, че $A_i \cap A_j = \emptyset$ при $i \neq j$

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n). \quad (2.1)$$

Следствие от аксиомите са свойствата:

- 1) $P(\bar{A}) = 1 - P(A)$
- 2) $P(\emptyset) = 0$
- 3) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(2.2)

Доказателство. 1) Събитията A и $\bar{A}$ са несъвместими като $A \cup \bar{A} = \Omega$ и аксиоми 2 и 3 $P(A \cup \bar{A}) = P(A) + P(\bar{A}) = 1$, откъдето следва 1).

2) $\bar{\Omega} = \emptyset \Rightarrow P(\emptyset) = P(\bar{\Omega}) = 1 - P(\Omega) = 1 - 1 = 0$.

3) Представяме $A \cup B$ като сума от несъвместими събития.

$$A \cup B = (A \cap B) \cup (A \cap \bar{B}) \cup (\bar{A} \cap B)$$

Следователно, $P(A \cup B) = P(A \cap B) + P(A \cap \bar{B}) + P(\bar{A} \cap B)$

Аналогично, $P(A) = P(A \cap B) + P(A \cap \bar{B})$, $P(B) = P(A \cap B) + P(\bar{A} \cap B)$. Тогава

$$P(A \cup B) = P(A \cap B) + P(A \cap \bar{B}) + P(\bar{A} \cap B) + P(A \cap B) - P(A \cap B) = P(A) + P(B) - P(A \cap B) \quad \spadesuit$$

Разглежда се също и така наречената условна вероятност:

Вероятността на едно събитие A , при допълнителното условие, че е настъпило някакво събитие B означаваме с $P(A|B)$ (или $P_B(A)$) и наричаме условна. Ако $P(B) > 0$ и $P(A \cap B)$ е вероятността за съвместното им появяване (вероятност на произведение от събития), то

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \quad (2.3)$$

От (2.3) следват следните формули за вероятност на произведение и сума на събития:

Вероятност на произведение (съвместна поява на събития):

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B).$$

- За n събития: $P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2|A_1) \dots P(A_n|A_1 \cap \dots \cap A_{n-1})$.
- За независими събития: $P(A \cap B \cap C) = P(A)P(B)P(C) \dots$

Формула (2.1) изразява, че вероятността за сума на n несъвместими събития е сума от вероятностите им.

Вероятността за сума на две съвместими събития се пресмята по формула (2.2), чието обобщение за n събития, не е удобно. За да намерим вероятността $P(A_1 \cup A_2 \dots \cup A_n)$ в този случай се постъпва по следния начин. Противоположното на събитието $C = A_1 \cup \dots \cup A_n$ - поява на поне едно от събитията $A_1, \dots, A_n$, е събитието $\bar{C} = \{\text{не е настъпило нито едно от събитията } A_1, \dots, A_n\}$, т.е. $\bar{C} = \bar{A}_1 \cap \dots \cap \bar{A}_n$. Тогава $P(C) = 1 - P(\bar{C})$, откъдето получаваме, че

Вероятността на сума (поява на поне едно от събитията) е

$$P(A_1 \cup \dots \cup A_n) = 1 - P(\bar{A}_1 \cap \dots \cap \bar{A}_n)$$

Пример 2.4. Гражданин търси на пазара 2 книги. Известно е, че вероятността да намери първата книга е равна на 0,3, да намери втората книга е 0,2, а да намери и двете е 0,04. Да се намери вероятността гражданинът да намери поне една от книгите.

Решение. Ако въведем събитията $A=\{\text{гражданинът намира първата книга}\}$, $B=\{\text{гражданинът намира втората книга}\}$, то $P(A)=0,3$, $P(B)=0,2$. Тогава $A \cap B = \{\text{гражданинът намира и двете книги}\}$ и $P(A \cap B)=0,04$. Търсим вероятността на събитието $C=\{\text{гражданинът намира поне една от книгите}\}$, т.е. $P(C)=P(A \cup B)=P(A)+P(B)-P(A \cap B)=0,3+0,2-0,04=0,46$.

Пример 2.5. Ученик трябва да избере 2 от три избираеми дисциплини френски, история и география. Той би избрал френски с вероятност $5/8$, история също с вероятност $5/8$. Каква е вероятността да избере: а) география; б) география и история; в) френски и история.

Решение. Означаваме $F=\{\text{ученикът избира френски}\}$, $I=\{\text{ученикът избира история}\}$, $G=\{\text{ученикът избира география}\}$, като $P(F)=P(I)=\frac{5}{8}$.

Ученикът трябва да избере два от общо три предмета, затова и трите събития $F \cup I$, $I \cup G$ и $G \cup I$ са достоверни. По формулата за сума на съвместими събития имаме $P(F \cup I)=P(F)+P(I)-P(F \cap I)=1$,

$$\text{откъдето } P(F \cap I)=P(F)+P(I)-P(F \cup I)=\frac{5}{8}+\frac{5}{8}-1=\frac{1}{4}.$$

а) Събитието $\bar{G}=\{\text{не избира география}\}=\{\text{избира френски и история}\}$, т.е. $\bar{G}=F \cap I$. Тогава $P(G)=1-P(\bar{G})=1-P(F \cap I)=1-\frac{1}{4}=\frac{3}{4}$.

б) Вероятността $P(I \cap G)$ ще намерим като използваме, че и събитието $I \cup G$ е достоверно. В $P(I \cup G)=P(I)+P(G)-P(I \cap G)=1$.

заместваме $P(I)=\frac{5}{8}$ и $P(G)=\frac{3}{4}$, получаваме $P(I \cap G)=\frac{5}{8}+\frac{3}{4}-1=\frac{3}{8}$.

в) Аналогично,

$$P(F \cup G)=P(F)+P(G)-P(F \cap G)=1 \Rightarrow P(F \cap G)=\frac{5}{8}+\frac{3}{4}-1=\frac{3}{8}.$$

Забележка. Събитията $F \cap I$, $I \cap G$ и $F \cap G$ също образуват пълна група събития, което лесно може да се провери.

Упражнения.

1. Нека $P(A \cap B)=1/4$, $P(\bar{A})=1/2$ и $P(B)=1/2$. $P(A \cup B)=?$.

2. Нека $\Omega=\{A, B, C\}$ е пространството от елементарни събития, настъпващи при даден експеримент и $P(A)=1/2$, $P(B)=1/3$. Да се намерят вероятностите на всички случайни събития, свързани с този експеримент.

3. Пространството от елементарни събития на случаен експеримент е $\Omega=\{\omega_1, \dots, \omega_5\}$ със вероятности съответно 0,1, 0,1, 0,2, 0,4 и 0,2. Нека

$A=\{\omega_1, \omega_2, \omega_3\}$, $B=\{\omega_1, \omega_3, \omega_5\}$. Да се намерят вероятностите $P(A)$, $P(\bar{A})$, $P(B)$, $P(\bar{B})$, $P(A \cap B)$, $P(A \cup B)$.

4. Опитът се състои в подхвърляне на три правилни монети. Нека $A=\{\text{на първата монета се пада Л}\}$, $B=\{\text{поне на две от монетите се пада Л}\}$. Да се състави пространството от елементарни събития за дадения опит. Да се намерят вероятностите $P(A)$, $P(B)$, $P(A \cap B)$, $P(A \cup B)$, $P(\bar{A} \cup \bar{B})$, $P(A \cup \bar{B})$, $P(A \cap \bar{B})$.

§3. Статистическа, класическа и геометрична вероятност.

Аксиоми 1, 2 и 3 позволяват чрез вероятностите на елементарните събития да се намери вероятността на произволно събитие, свързано със случайния експеримент. Въпросът за това как да се определи вероятността на елементарните събития не се разглежда. На практика тези вероятности се определят в зависимост от условията на експеримента по начин, който се съгласува най-добре с практическите наблюдения.

Статистическа вероятност. Както отбелязахме, една интуитивна оценка на шанса за сбъждане на едно събитие е относителната му честота, получена при многократно повторение на експеримента. Важен експериментално доказан факт е устойчивостта на честотата. Забелязва се, че при увеличение на броя на опитите в серията честотата на събитието се колебае около една и съща стойност независимо нито от броя на опитите, нито от серията опити. Например, в няколко серии от по 20 изстрела стрелец е уличил целта (събитие A) съответно 5, 4, 3, 6, 7, 8 и т.н. пъти, но средния брой попаденията е 5, т.е. средната стойност на относителната честота е $1/4$. За този стрелец това число е относително постоянно и характеризира неговото умение.

Числото $P(A)$, към което се стремят относителните честоти

$$\frac{m_1}{n_1}, \frac{m_2}{n_2}, \frac{m_3}{n_3}, \dots, \frac{m_i}{n_i}, \dots$$

на поява на събитието A при увеличаване на броя n_i на опитите в сериите, се нарича статистическа вероятност на събитието A .

По този начин статистическата вероятност е вероятност, получена на базата на опитни данни. Такъв характер имат много събития, като заболяване от дадена болест, повреда на уред за определен срок, производство на некачествено изделие и др.

Пример 3.1. Изследвана е точността при изпълнение на сервиза на тенисист и са получени следните резултати:

Брой на опитите	100	500	1000
Точни попадения	76	370	738
Относит. честота	0,76	0,74	0,736

От тук можем да приемем, че вероятността за правилно изпълнение на сервиза на този спортист е 0,74.

Класическа вероятност. Както отбелязахме, за да установи, че падането на лице или герб при подхвърляне на една правилна монета са равновероятни, Буфон е направил 4040 опита (пример 2.1). Но можеше да се разсъждава така. Пространството от елементарните събития $\Omega = \{I, G\}$ при подхвърлянето на монета има два елемента. Събитието $A = \{\text{пада се Л}\}$ има един благоприятен случай. Тогава може да приемем, че $P(A) = \frac{1}{2}$, $P(\bar{A}) = \frac{1}{2}$, което напълно се покрива с опитните данни.

Ще отбележим, че важни условия за такива разсъждения са: 1) елементарните събития са краен брой и 2) те са равновероятни, т.е. поради симетричност нямаме основание да считаме, че някое елементарно събитие по-често се случва от другите. В такъв случай имаме основание за следната дефиниция за вероятност на събитие.

Нека множеството $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$, описващо резултатите от случайния експеримент, се състои от краен брой равновероятни изходи (елементарни събития) и нека A е събитие, което се наблюдава в този експеримент.

Вероятност на събитието A , ще наричаме отношението $P(A) = \frac{m}{n}$, където m и n са съответно броят на благоприятните за A и броят на всички възможни и равновероятни елементарни събития.

Пример 3.2. В кутия има 2 бели и три черни топки. От кутията се изтеглят 3 топки. Каква е вероятността и трите топки да са черни (събитие C), ако изтеглените топки: а) се връщат; б) не се връщат?

Решение. Ако $A_i = \{i\text{-тата топка е черна}\}$, то $C = A_1 \cap A_2 \cap A_3$.

а) Събитията A_i са зависими. Тогава

$$P(C) = P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) = \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{1}{3} = \frac{1}{10}.$$

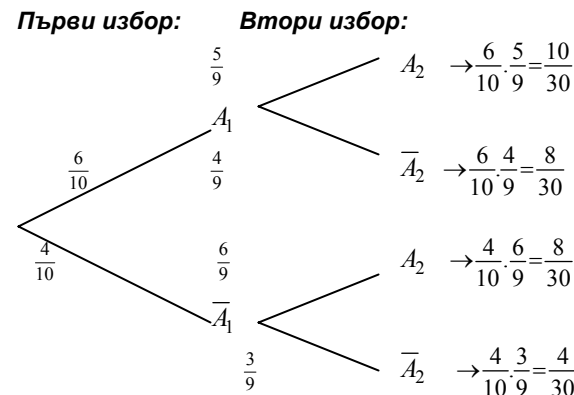
б) Тъй като изтеглените топки се връщат, то A_i са независими и

$$P(C) = P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3) = \frac{3}{5} \cdot \frac{3}{5} \cdot \frac{3}{5} = \frac{27}{125}.$$

Пример. 3.3. Кутия съдържа 6 бели и 4 черни топки. От кутията се изваждат две топки. Да се намери вероятността на събитията: а) $A = \{\text{двете топки да са бели}\}$; б) $B = \{\text{двете топки са различен цвят}\}$; в) втората топка да е бяла.

Решение: Означаваме събитията: $A_i = \{i\text{-тата изтеглена топка е бяла}\}$, $i=1,2$. Ще решим задачата с помощта на следната схема (граф със структура на дърво, за което върховете са настъпилите събития, а по дъгите са написани вероятностите им)

При първия избор в кутията има 10 топки, от които 6 са бели, следователно, имаме $P(A_1) = \frac{6}{10}$, $P(\bar{A}_1) = 1 - \frac{6}{10} = \frac{4}{10}$, които са нанесени по дъгите. Ако първата изтеглена топка е била бяла (настъпило е събитието A_1), то вероятността втората изтеглена топка да е бяла (събитие A_2) е равна на $P(A_2|A_1) = \frac{5}{9}$, защото от останалите 9 топки 5 са белите.



По същия начин са получени числата и по останалите дъги (също условни вероятности). Тогава:

а) $P(A) = P(A_1 \cap A_2) = P(A_1) \cdot P(A_2|A_1) = \frac{6}{10} \cdot \frac{5}{9} = \frac{10}{30}$ (умножават се вероятностите по пътя от началото до върха A_2). Тук сме приложили формулата за вероятност на произведение от събития).

б) Тъй като $B = (A_1 \cap \bar{A}_2) \cup (\bar{A}_1 \cap A_2)$, то по формулата за сума от несъвместими събития получаваме

$$P(B) = P(A_1 \cap \bar{A}_2) + P(\bar{A}_1 \cap A_2) = \frac{8}{30} + \frac{8}{30} = \frac{8}{15}.$$

в) Търсим вероятността за събитието A_2 . Случаите, при които настъпва събитието са: $A_1 \cap A_2$, вероятността за което е $P(A_1) \cdot P(A_2|A_1) = \frac{6}{10} \cdot \frac{5}{9} = \frac{1}{3}$ и $\bar{A}_1 \cap A_2$, вероятността за което е $P(\bar{A}_1) \cdot P(A_2|\bar{A}_1) = \frac{4}{10} \cdot \frac{6}{9} = \frac{2}{9}$.

Следователно,

$$P(A_2) = P(A_1)P(A_2|A_1) + P(\bar{A}_1)P(A_2|\bar{A}_1) = 5/9$$

(виж теорема 5.1 за пълната вероятност).

Геометрична дефиниция за вероятност. Да разгледаме опит, който може да се интерпретира като случайно попадение на точка M в геометрична област Ω – отсечка, равнинна фигура или пространствено тяло с крайна мярка S_Ω (т.е. съответно с крайна дължина, лице или обем)

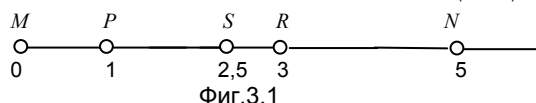
Ако условията на опита са такива, че вероятността точката M да попадне в област $D \subseteq \Omega$ (събитие $A = \{M \in D\}$) е пропорционална на мярката S_D (дължина, лице или обем) и не зависи от местоположението на D в Ω , то съгласувано с практическите резултати е геометричната вероятност

$$P(A) = \frac{S_D}{S_\Omega},$$

където S_Ω е мярката на областта Ω .

Пример 3.3. От отсечката MN с дължина 5 cm по случаен начин е избрана точката Q . Каква е вероятността: а) отсечката MQ да е по-къса от отсечката QN (събитие A); б) точката да се намира на разстояние не повече от 1 cm от M и на не повече от 2 cm от N (събитие B).

Решение. а) Ако S е среда на отсечката MN , то събитието A означава, че $Q \in MS$, следователно, $P(A) = P(Q \in MS) = \frac{|MS|}{|MN|} = \frac{1}{2}$.



б) От фиг. 3.1. получаваме

$$P(B) = P(Q \in MP) + P(Q \in RN) = \frac{|MP|}{|MN|} + \frac{|RN|}{|MN|} = \frac{1}{5} + \frac{2}{5} = \frac{3}{5}.$$

За разглежданите три начина за определяне на вероятност на събития може да се провери, че са изпълнени всички аксиоми за вероятност на събитие.

Упражнения.

1. От отсечка AB с дължина 12 cm произволно се посочва точка M при това вероятността точката да попадне в който и да е интервал е пропорционална на дължината му. Каква е вероятността за това лицето на квадрата със страна AM да е по-голямо от 36 cm² и по-малко от 81 cm².

§4. Елементи от комбинаториката.

В случая на краен брой равновероятни изходи от опита (класическа вероятност) за намирането на вероятността на едно събитие не е необходимо да се правят реални опити. Достатъчно е да се пресметнат

благоприятните и всички възможни случаи. За последното се налага използването на някои формули от комбинаториката, които ще изложим в резюме.

Нека е дадено множеството $M = \{e_1, e_2, \dots, e_n\}$ от n елемента. Разглеждаме различните възможности за групиране на елементите на M по k елемента. Най-общо, групите (в комбинаториката те се наричат съединения) можем да разделим на ненаредени (комбинации) и наредени (вариации):

Комбинация от n елемента k -ти клас се нарича всяка ненаредена група от k елемента на множеството (т.е. групите се различават една от друга по състава на елементите, но не и по подреждането им в групата).

Вариация от n елемента k -ти клас се нарича всяка наредена група от k елемента на множеството (т.е. групите се различават една от друга както по състава на елементите, така и по подреждането им в групата).

В някои случаи е възможно в групата един елемент да участва повече от един път (група с повторение), а в други това не е възможно (група без повторение).

Броят на различните видове съединения се изчислява по формулите:

Комбинации от n елемента k -ти клас

а) без повторения $C_n^k = \binom{n}{k} = \frac{n(n-1)(n-2)\dots(n-k+1)}{k!} = \frac{n!}{k!(n-k)!}, \quad (4.1)$

б) с повторения $\tilde{C}_n^k = \binom{n+k-1}{k}.$

Вариации от n елемента k -ти клас

а) без повторения $V_n^k = n(n-1)(n-2)\dots(n-k+1)$

б) с повторения $\tilde{V}_n^k = n^k.$

Частен случай на вариация без повторение е пермутацията – всяко размятане на елементите в множеството.

Броят на пермутациите на множество от n елемента е $P_n = V_n^n = n!.$

Ще отбележим, че по дефиниция, $C_n^0 = 1, V_n^0 = 1, P_0 = 1$.

Пример 4.1. Каква е вероятността произволно избрано трицифрено число да се състои от различни цифри.

Решение. Всяко тризначно число (например числото 212) е наредена тройка от десетте цифри 0,1,2,...,9, като е възможно и повторение на една или повече цифри, т.е. броят на всички числа, записани с три цифри е $\tilde{V}_{10}^3 = 10^3 = 1000$. Но тук са включени двузначните

и еднозначните числа, броят на които е $\tilde{V}_{10}^2 = 10^2 = 100$. Следователно, броят на всички тризначни числа (броят на всички елементарни изходи) е $n = 1000 - 100 = 900$. Броят на числата, записани с различни цифри е $m = V_{10}^3 - V_9^2 = 10 \cdot 9 \cdot 8 - 9 \cdot 8 = 648$. Следователно, $P(A) = \frac{648}{900} = 0,72$. ♦

Пример 4.2. При хвърляне на две зарчета се отчита сумата S от падналите точки. Установено е, че по-често се случва, например, сумата да е 7, отколкото 6, и че най-рядко се пада сума 2 или 12. Какъв модел за пространство на елементарните събития трябва да се избере в този опит.

Решение. 1) Тъй като се интересуваме само от сумата от падналите точки, то естествено е да въведем 11 елементарни събития: сумата S да е равна на 2, 3, ..., 11, 12. Но тези елементарни събития не са равновероятни, както не отговаря на наблюденията.

2) Нека $\omega_{ij} = \{\text{на първото зарче са падат } i \text{ точки, а на второто } - j \text{ точки}\}$. Ако зарчетата са еднакви, то те са неразличими за нас и резултатът ω_{ij} не се различава от резултата ω_{ji} , т.е. може да приемем, че всеки изход от едно хвърляне е комбинация от 6 елемента втори клас с повторения. Следователно, броят на всички изходи е $n = \tilde{C}_6^2 = \frac{7 \cdot 6}{2} = 21$,

а събитията $\{S=6\}$ и $\{S=7\}$ настъпват в равен брой случаи ($6=1+5=2+4=3+3$, $7=1+6=2+5=3+4$). Тогава вероятностите тези две събития би трябвало да саравни на $3/21$, което показва, че и този модел не е добър.

3) Ако приемем, че зарчетата са различни (например са различни по цвят), то пространството от елементарните изходи е $\Omega = \{\omega_{ij}\}$, елементарните изходи ω_{ij} ($i, j=1, \dots, 6$) са равновероятни и броят им е

$n = \tilde{V}_6^2 = 6^2 = 36$. В този случай $P(S=6) = \frac{5}{36}$, а $P(S=7) = \frac{6}{36}$, което

съответства на наблюденията ни.

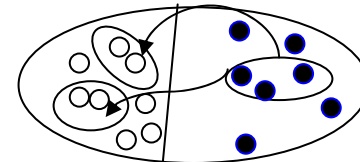
Забележка 4.1. Пример 4.2 показва, че при определяне на пространството от елементарни събития използването на вариации с повторение е за предпочитане пред комбинации с повторение, независимо дали практически може да се различи наредбата в групите.

Пример 4.3. В кутия има M бели и $N-M$ черни топки. От кутията се изтеглят без връщане k топки ($k \leq \min(M, N-M)$). Разглеждаме събитието A - точно r от изтеглените топки да са бели. Да се докаже, че

$$P(A) = \frac{C_M^r C_{N-M}^{k-r}}{C_N^k} \quad (4.2)$$

Решение. Всеки k изтеглени топки са една комбинация от n елемента k -ти клас, следователно, всички възможни изходи са $n = C_N^k$.

Благоприятен случай е всяка група, в която имаме r бели и $k-r$ черни топки. За да ги преброим, разделяме множеството на топките на две непресичащи се множества: множеството на белите топки (на брой M) и множеството на черните топки (общо $N-M$ топки). Тогава за фиксирани r бели топки, благоприятен е всеки набор от $k-r$



черни топки, броят на които е C_{N-M}^{k-r} . Но общият брой на r -торките бели топки е C_M^r . Следователно благоприятните случаи са $m = C_M^r \cdot C_{N-M}^{k-r}$, откъдето следва формула (4.2).

Упражнения.

1. За кои пространства от елементарни събития е целесъобразно да се избере класическата вероятност.
2. Зарче се хвърля 3 пъти. Да се намери вероятността на трите зарчета да се падне четен брой точки.
3. В библиотека има книги, разпределени в 10 раздела. Постъпили са 4 поръчки за литература. Считайки, че е възможен произволен състав на поръчаната литература, да се наметат вероятностите за събитията $A = \{\text{книгите са от един и същ раздел}\}$, $B = \{\text{книгите са от различни раздели}\}$, $C = \{\text{две от книгите са от един и същ раздел, а останалите две - от различни раздели}\}$.
4. Да се намери вероятността при случайно подреждане на 10 книги три определени книги да се намират една до друга.
5. На пет картички са написани буквите А, И, С, Ф, Я. Картичките се разполагат в ред по произволен начин. Каква е вероятността да се получи думата СОФИЯ?
6. Телефонен номер се състои от 6 цифри. Да се намери вероятността: а) всички цифри да са различни, б) номерът да завършва на 240.
7. На 10 картички са написани буквите А, А, И, И, К, С, С, Т, Т, Т. Картичките се разполагат в ред по произволен начин. Каква е вероятността да се получи думата СТАТИСТИКА?
8. В кутия има 4 бели и 2 черни топчета. Изтеглят се 2 топчета. Каква е вероятността топчетата да са различен цвят?
9. В кутия има 10 топчета. Вероятността да се изтеглят 2 бели топчета е $2/15$. Колко бели топчета има в кутията?
10. В една кутия има a бели и b черни топки, а в друга - c бели и d черни. От двете кутии се изтеглят по една топка. Да се намери вероятността изтеглените топки: а) да се еднакъв цвят; б) да се различен цвят.
11. В група от 40 студенти има 10 спортисти. По случаен начин са избрани 20 студента. Да се намери вероятността между тях да има 5 спортисти.
12. В кутия има n бели и m черни топки. Изтеглят се k ($k > m$) топки. Каква е вероятността в кутията да са останали само бели топки.
13. В кутия има 4 бели, 4 червени и 4 сини топчета. Каква е вероятността от 4 изтеглени топчета 1 да е бяло, 2 червени и 1 синьо.

§5. Формула за пълната вероятност. Формули на Бейс.

В пример 3.2.в) намерихме безусловната вероятност на събитието $A_2 = \{\text{втората изтеглена топка е бяла}\}$ с помощта на условните му вероятности. Получената формула е частен случай на теоремата:

Теорема 5.1. (формула за пълната вероятност). Нека A е произволно събитие и $H_1, H_2, \dots, H_n$ е пълна група несъвместими събития, за които $P(H_i) > 0$ ($i=1, \dots, n$). Тогава

$$P(A) = \sum_{i=1}^n P(H_i)P(A|H_i) \quad (5.1)$$

Доказателство. От условието на теоремата следва, че при $i \neq j$ $H_i \cap H_j = \emptyset$ и $H_1 \cup \dots \cup H_n = \Omega$.

Формула (5.1) следва от представянето

$$A = A \cap \Omega = A \cap (H_1 \cup H_2 \cup \dots \cup H_n) = (A \cap H_1) \cup (A \cap H_2) \cup \dots \cup (A \cap H_n),$$

В последната сума събитията са несъвместими. Прилагаме формулите за сума, а след това за произведение на вероятностите и получаваме

$$\begin{aligned} P(A) &= P(A \cap H_1) + P(A \cap H_2) + \dots + P(A \cap H_n) = \\ &= P(H_1)P(A|H_1) + P(H_2)P(A|H_2) + \dots + P(H_n)P(A|H_n). \end{aligned}$$

Тази теорема се прилага, когато се търси безусловната вероятност на събитие, за което са известни условните му вероятности при всички възможни случаи (образуващи пълна група несъвместими събития).

Пример 5.1. Група от 20 спортиста се състои от 7 начинаещи, 10 добре и 3 отлично подготвени. Вероятността, че начинаещ спортист ще се справи с дадено упражнение е 0,4, а добре и отлично подготвен - съответно 0,8 и 1. Каква е вероятността случайно извикан спортист да изпълни упражнението (събитие A)?

Решение. Извиканият спортист може да бъде начинаещ (събитие H_1) добре подготвен (събитие H_2) или отлично подготвен (събитие H_3). По условие $P(H_1) = \frac{7}{20}$, $P(H_2) = \frac{10}{20}$, $P(H_3) = \frac{3}{20}$ (пълна група несъвместими събития).

Вероятността спортист от дадена група да изпълни упражнението е

$$P(A|H_1) = 0,4, \quad P(A|H_2) = 0,8, \quad P(A|H_3) = 0,9.$$

Тогава по формула (5.1)

$$\begin{aligned} P(A) &= P(H_1)P(A|H_1) + P(H_2)P(A|H_2) + P(H_3)P(A|H_3) = \\ &= \frac{1}{20}(7 \cdot 0,4 + 10 \cdot 0,8 + 3 \cdot 1) = 0,69. \end{aligned}$$

Забележка. Задачата може да се реши и с помощта на схемата, приложена в пример 3.2.

Теорема 5.2. (формули на Бейс). Нека A е произволно събитие и $H_1, H_2, \dots, H_n$ е пълна система несъвместими събития, за които $P(H_i) > 0$ ($i=1, \dots, n$). Тогава

$$P(H_k|A) = \frac{P(H_k)P(A|H_k)}{\sum_{i=1}^n P(H_i)P(A|H_i)}, \quad k=1, 2, \dots, n. \quad (5.2)$$

Доказателство. Имаме $P(H_k|A) = \frac{P(A \cap H_k)}{P(A)} = \frac{P(H_k)P(A|H_k)}{\sum_{i=1}^n P(H_i)P(A|H_i)}$. ♦

Пример 5.2. При условията от предната задача извиканият спортист е изпълнил упражнението. Каква е вероятността той да е бил начинаещ?

Решение. Тук се търси условната вероятност за събитието H_1 , ако е настъпило събитието A , за която получаваме

$$P(H_1|A) = \frac{P(H_1)P(A|H_1)}{P(H_1)P(A|H_1) + P(H_2)P(A|H_2) + P(H_3)P(A|H_3)} = \frac{0,14}{0,66} \approx 0,20.$$

Както се вижда, $P(H_1|A) < P(H_1)$, което би трябвало да се очаква. ♦

Формулите на Бейс намират приложение, когато за настъпването на дадено събитие съдим по косвени признаци. Такива случаи имаме в много области от медицината, физиката, астрономията, при които вероятността за верността на дадено предположение определяме по настъпването на друго събитие, което сме в състояние да наблюдаваме.

Пример 5.3. (Диагностичен тест) Вероятността жителите на дадена област да се разболеят от дадено заболяване е 0,15. Болестта се открива с тест, който дава положителен резултат с вероятност 0,9 при наличие на заболяване и с вероятност 0,02 при отсъствие на болестта. Да се намери вероятността гражданин от областта, за който тестът е положителен, да е болен от тази болест.

Решение. Събитието $H = \{\text{гражданинът страда от болестта}\}$ и неговото отрицание $\bar{H}$ образуват пълна група, $P(H) = 0,15$, $P(\bar{H}) = 0,85$. Искаме да преоценим тези вероятности след като е настъпило събитието $A = \{\text{тестът за гражданина е положителен}\}$ като знаем, че $P(A|H) = 0,9$ и $P(A|\bar{H}) = 0,02$. По формула (5.2)

$$P(H|A) = \frac{P(H)P(A|H)}{P(H)P(A|H) + P(\bar{H})P(A|\bar{H})} = \frac{0,15 \cdot 0,9}{0,15 \cdot 0,9 + 0,85 \cdot 0,02} = 0,89. \quad \blacklozenge$$

Упражнения.

1. В група от 20 ловци има 4 отлични, 10 добри и 6 посредствени стрелци като вероятността при един изстрел да улови целта е съответно 0,9 за отличен стрелец, 0,7 за добър и 0,5 за посредствен. Двама от стрелците са произвели по един изстрел. Каква е вероятността да има две попадения.

2. Известно е, че 96% от продукцията на завод е стандартна. Поради опростена система за контрол стандартно изделие се признава за стандартно с вероятност 0,98, а нестандартно се признава за стандартно с вероятност 0,05. а) Да се определи вероятността за това, че след проверката стандартно изделие ще бъде признато за стандартно.

ВЕРОЯТНОСТ НА СЪБИТИЕ – ОСНОВНИ РЕЗУЛТАТИ

1) Случайни събития се наблюдават в експерименти, резултатът от които не е еднозначно определен (случаен, стохастичен експеримент).

2) Математичен модел на случайните експерименти е пространството Ω от елементарните събития. Случайните събития се разглеждат като подмножества на Ω , като подмножеството се състои от всички елементарни събития, с настъпването на които настъпва и случайното събитие (благоприятни случаи). С Ω се означава достоверното, а с $\emptyset$ - невъзможното събитие.

3) За две (и повече) събития A и B се въвеждат действия събиране $A \cup B$ и умножение $A \cap B$, съответстващи и притежаващи свойствата на действията обединение и сечение на множества.

4) Пълното описание на едно случайно събитие A включва: 1) съвкупността от всички благоприятни случаи; 2) вероятността за настъпване на събитието - мярка за честотата на събъждане на събитието със свойствата:

$$1) 0 \leq P(A) \leq 1; \quad 2) P(\Omega) = 1 \Rightarrow P(\emptyset) = 0;$$

$$3) \text{ Ако } A_i \cap A_j = \emptyset \text{ при } i \neq j, \text{ то } P\left(\bigcup_n A_k\right) = \sum_n P(A_k).$$

5) Вероятността е число, което се определя в зависимост от условията на експеримента. Начини за получаване:

• статистически (базиран на опитни данни) – статистическа вероятност.

• теоретичен (базиран на математичен модел на експеримента) – такива са, например, класическата или геометричната вероятност.

6) Основни взаимовръзки между две събития:

• $A \subseteq B$ (A влече B): $P(A) \leq P(B)$,

• $A = B$ (A еквивалентно на B) $P(A) = P(B)$;

• A и B са несъвместими, ако $P(A \cap B) = 0$.

• A и B са независими – ако условните вероятности $P(A|B)$ и $P(B|A)$ са равни съответно на безусловните вероятности $P(A)$ и $P(B)$, т.е.

$$P(A|B) = P(A) \text{ и } P(B|A) = P(B).$$

В противен случай, събитията се наричат зависими.

7) Формули за пресмятане на вероятност на случайно събитие:

1. Вероятност на произведение (съвместна поява на събития):

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B).$$

• За n събития: $P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2|A_1) \dots P(A_n|A_1 \cap \dots \cap A_{n-1})$.

• За независими събития: $P(A \cap B) = P(A)P(B)$.

2. Вероятност на сума (поява на поне едно от събитията)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

• За n събития е по-удобна формулата

$$P(A_1 \cup \dots \cup A_n) = 1 - P(\bar{A}_1 \cap \dots \cap \bar{A}_n).$$

• За несъвместими събития: $P(A \cap B \cap C) = P(A) + P(B) + P(C)$.

3. Ако $H_1, H_2, \dots, H_n$ образуват пълна група несъвместими събития и A е произволно събитие, то:

• $P(A) = \sum_{i=1}^n P(H_i)P(A|H_i)$ (формула за пълната вероятност)

• $P(H_k|A) = \frac{P(H_k)P(A|H_k)}{\sum_{i=1}^n P(H_i)P(A|H_i)}$, $k=1,2,\dots,n$ (формули на Бейс).

Общи задачи.

1. Зарче се хвърля 2 пъти. Да се опише пространството от елементарни събития Ω . Каква е вероятността за събитията: $A=\{\text{падат се различни точки}\}$; $B=\{\text{сумата от точките е } 3\}$; $C=\{\text{сумата от точките е } 5\}$; $D=\{\text{сумата от точките е } 6\}$; $F=\{\text{сумата от точките е четно число като на второто се пада } 6\}$.

2. На пет картички са написани буквите Е, О, Т, Р, Р. Изтеглят се последователно:

а) 5 картички. Каква е вероятността да се получи думата ТЕРОР.

б) 3 картички. Каква е вероятността да се получи думата ТОР.

3. Избира се случайно едно тризначно число. Каква е вероятността две от цифрите му да са еднакви,

а) ако в десетичния запис на числото няма нула,

б) в общия случай.

4. На картички са написани цифрите 1, 2, 3, 4 и 5. Ако картичките са подредени по случаен начин каква е вероятността да се е получило четно число.

5. В кутия има 5 бели и 5 черни топчета. Последователно се изтеглят всички топчета. Каква е вероятността цветовете им да се редуват?

6. На скамейка седят по случаен начин 5 души. Каква е вероятността 3 определени лица да бъдат един до друг.

7. В автобус има 5 пътника. Автобусът спира на 8 спирки. Каква е вероятността пътниците да спазват на различни спирки?

8. Система се състои от 5 компонента, означени с 1, 2, 3, 4 и 5. Всеки компонент може да работи (с означение 1) или да не работи (с означение 0). Да се дефинира

пространството от елементарните изходи и да се изрази като негово подмножество събитието $A = \{\text{броят на работещите компоненти е по-голям от броя на неработещите}\}$.

9. Колода от 52 разбъркани карти е разделена на 2 половини. Да се намерят вероятностите на събитията: $A = \{\text{във всяка половина има точно 2 аса}\}$. $B = \{\text{в една от половините няма нито едно асо}\}$

10. Избрано е случайно едно число. Ако $A = \{\text{числото завършва на нула}\}$, $B = \{\text{числото се дели на 5}\}$, какво означават събитията $B \setminus A$, $A \cap B$, $A \cup B$.

11. Нека A , B и C са три произволни събития. Да се изразят чрез тези събития събитията:

- а) настъпили са трите събития; б) настъпило е поне едно от събитията;
- в) настъпило е едно събитие; г) не е настъпило нито едно събитие;
- д) настъпили са не повече от две събития

12. Хвърлени са 6 правилни зарчета. Каква е вероятността за падането на: а) поне една; б) точно една; в) точно две единици.

13. В лотария вероятността за печалба е 0,2. Да се определи колко билета най-малко трябва да се закупят, че вероятността да се получи поне една печалба да бъде не по-малка от

- а) 0,65; б) 0,9; в) 0,99.

14. Да се намери вероятността при случайно подреждане на 10 книги три определени книги да се намират една до друга.

15. В кутия има 2 бели и 3 черни и 5 сини топчета. Изтеглят се 3 топчета. Каква е вероятността топчетата да са различен цвят.

16. Зарче е направено така, че 1 точка се пада 2 пъти по-често от останалите точки, вероятността на които е една и съща. Зарчето се хвърля 2 пъти.

а) Изпълнени ли са условията за прилагане на класическата дефиниция за вероятност в този опит.

б) Да се намери вероятността на събитията $A = \{\text{броят на точките при второто хвърляне са повече от броя на точките от първото}\}$, $B = \{\text{сборът от точките е равен на 7}\}$, техните противоположни събития и на събитията $A \cap B$ и $A \cup B$.

17. За да стигне до работното си място, гражданин може да ползва автобус или трамвай, като честотата на трамваите е два пъти по-голяма от честотата на автобусите. Ако вземе автобус, вероятността да закъснее е 0,2, ако вземе трамвай – 0,3.

а) Да се намери вероятността гражданинът да закъснее.

б) Известно е, че гражданинът е закъснял. Каква е вероятността този ден той да е пристигнал със трамвай.

в) Каква е вероятността през една работна седмица (5 дни) гражданинът да закъснее два пъти.

г) Ако през работната седмица има поне едно закъснение, служителът бива глобен. Каква е вероятността гражданинът да бъде глобен за закъснение 1) през дадена седмица; 2) през даден месец.

18. Гражданин е закупил три билета от лотария: по един от 1, 2 и 5 лева, вероятността за печалба от които е съответно 1/10, 1/5 и 1/4. Да се намерят вероятностите на събитията: $A = \{\text{печелят билетите от 1 и 5 лв}\}$, $B = \{\text{печели само един билет}\}$, $C = \{\text{печели поне един билет}\}$.

19. От 5 студента, явяващи се на изпит, двама знаят 10 въпроса от 30, двама – 20 от 30, а един – всички въпроси. Изпитът е взет, ако студентът отговори на един случайно изтеглен въпрос.

а) Намерете вероятността произволно извикан студент да вземе изпита.

б) Извиканият студент е издържал изпита. Каква е вероятността той да е чел само по 10 въпроса от конспекта.

20. Колко прави могат да се прекарат: а) през 8 точки, никои три от които не лежат на една права; б) през 10 точки, четири от които лежат на една права?

21. Едно изделие се обработва на една от две машини M_1 и M_2 . Вероятността за повреда на машините M_1 и M_2 са съответно 0,4 и 0,6. Известно е също, че 90% от обработената на машина M_1 и 70% от обработената на машина на M_2 продукция е първокачествена. Да се пресметне:

а) вероятността едно изделие да е първо качество.

б) вероятността изделието да е обработено на машината M_1 , ако се знае, че то е не е първо качество.

22. В две еднакви кутии има топки, които се различават само по цвят. В първата кутия има 5 бели, 2 черни и 3 червени топки, а във втората – 3 бели, 1 черна и 6 червени топки.

а) От кутията е извадена по една топка. Каква е вероятността двете топки да имат един и същ цвят?

б) Извадена е топка от една от кутиите. Каква е вероятността топката да е бяла?

в) Изтеглена е една черна топка. Каква е вероятността топката да е била в първата кутия?

23. Една монета се подхвърля 3 пъти. Да се опише пространството Ω на елементарните събития и да се изразят като негови подмножества събитията $A = \{\text{при първото подхвърляне се пада лице}\}$, $B = \{\text{при второто подхвърляне се пада лице}\}$, $C = \{\text{пада се повече лице, отколкото герб}\}$ и да се намерят вероятностите им, ако се предполага, че всички елементарни събития имат равни вероятности.

24. В 75% от случаите клиентите на интернет-център се обаждат по телефона за да се оплачат от качеството на услугата, в 5% – да поискат справка или информация, а останалите обаждания са свързани с персонала на сервиза. От оплакванията 77% се дължат на неизправности на връзката, 20% – на непълноти в софтуера, а останалите 3% – на неправилно инсталиране на софтуера. Да се определи вероятността за събитията:

а) $A = \{\text{при обаждане по телефона клиент да се оплаче от качеството на връзката}\}$.

б) $B = \{\text{От три обаждания две да са справки и едно оплакване за качеството на връзката}\}$.

25. В две еднакви кутии има топки, които се различават само по цвят. В първата кутия има 5 бели, 2 черни и 3 червени топки, а във втората – 3 бели, 1 черна и 6 червени топки. От кутията е извадена по една топка. Каква е вероятността двете топки да имат един и същ цвят.

СЛУЧАЙНИ ВЕЛИЧИНИ

§6. Случайна величина. Видове. Функция на разпределение на случайна величина.

Случайни са тези величини, които приемат случайни стойности, т.е. това са величини, които се измерват в случайните експерименти.

Например:

- опит – 5 изстрела по мишена, случайна величина – брой на попаденията;
- опит – експлоатация на уред, случайна величина – продължителност на безотказна работа;
- опит – измерване на физична величина, случайна величина – грешка на измерването.

Както се вижда, често пъти резултатите от един случаен експеримент са числа. В други експерименти те не са числа, но на тях могат да бъдат съпоставени числени стойности, с помощта на които да се изследват случайните събития. Каква стойност ще приеме случайната величина, се определя едва след като е известен изходът от случайния експеримент, т.е. случайната величина е функция на изходите от експеримента.

Случайна величина (случайна променлива) се нарича всяка еднозначна реална функция на елементарните случаи ω , описващи даден случаен експеримент, т.е. $\xi = \xi(\omega)$, $\omega \in \Omega$.

Пример 6.1. При подхвърлянето на една монета два пъти елементарни събития са ЛЛ, ЛГ, ГЛ, ГГ. Ако с ξ означим броя на поява на лицето на монетата, то $\xi(ЛЛ) = 2$, $\xi(ЛГ) = \xi(ГЛ) = 1$, $\xi(ГГ) = 0$. ♦

Случайните величини ще означаваме с гръцки букви ξ , η , ζ и т.н. Ако направим конкретно измерване на случайната величина, ще получим конкретна стойност x , която се нарича реализация на ξ и е една от възможните ѝ стойности.

Случайните величини, които ще разглеждаме, се делят по отношение на множеството от възможните си стойности на:

Дискретни - случайни величини, множеството от възможните стойности на които е крайно или изброимо много.
Непрекъснати - случайни величини, множеството от възможните стойности на които е интервал, краен или безкраен.

Ще отбележим че има също и сингулярни случайни величини – величини, които не са от горните два вида.

За да е определена една случайна величина не е достатъчно да знаем какво е множеството от възможните ѝ стойности. Трябва да знаем също каква е вероятността случайната величина ξ да приеме конкретна възможна стойност (например $\xi = a$) или стойност в дадено числово множество (например $a \leq \xi < b$, $\xi < c$, множеството от четни числа), т.е.

за да познаваме една случайна величина трябва да знаем как да изчисляваме вероятността на произволно събитие, отнасящо се до случайната величина.

Правилото, с помощта на което може да бъде намерена вероятността на произволно събитие, свързано със случайната величина наричаме закон за разпределението на вероятностите на величината.

Всяка случайна величина има свой закон на разпределение.

Един начин за определянето му е за всяка реална стойност на числото x да бъде дадена вероятността на събитието $\{\xi < x\}$

Функцията $F_{\xi}(x) = P(\xi < x)$, определена за всяко $x \in (-\infty, \infty)$ се нарича функция на разпределение на случайната величина ξ .

Свойства на функцията на разпределение:

- 1) $0 \leq F_{\xi}(x) \leq 1$ за всяко x ,
- 2) $F_{\xi}(x)$ е ненамаляваща за всяко x
- 3) $\lim_{x \rightarrow -\infty} F_{\xi}(x) = 0$, $\lim_{x \rightarrow \infty} F_{\xi}(x) = 1$;

Доказателството на тези свойства произтича от факта, че за всяко x функцията $F_{\xi}(x)$ е равна на вероятността на събитието $\xi < x$.

Ако за една случайна величина ξ функцията $F_{\xi}(x)$ е дадена, то можем да намерим вероятността на всяко събитие от вида $\{a \leq \xi < b\}$.

Действително, да разгледаме събитията $\{\xi < b\}$ и $\{\xi < a\}$, за които, съгласно дефиницията на функцията на разпределение имаме $P(\xi < b) = F_{\xi}(b)$, $P(\xi < a) = F_{\xi}(a)$. Очевидно, $P(\xi < b) = P(\xi < a) + P(a \leq \xi < b)$, откъдето

$$P(a \leq \xi < b) = F_{\xi}(b) - F_{\xi}(a). \quad (6.1)$$

Забележка 6.1. Вероятността $P(\xi = a)$ може да изчислим от формула (6.1) при граничния преход $b \rightarrow a$, т.е.

$$P(\xi = a) = \lim_{b \rightarrow a, b > a} [F_{\xi}(b) - F_{\xi}(a)] \quad (6.2)$$

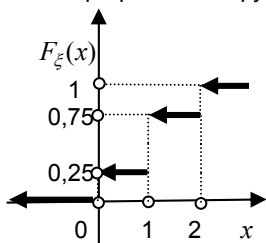
Пример 6.2. Да се намери функцията на разпределение на величината от пример 6.1.

Решение. Възможните стойности на величината са 0, 1 и 2 и затова трябва да разгледаме случаите:

1. $x \leq 0$. Най-малката стойност на ξ е 0, затова събитието $\xi < x \leq 0$ е невъзможно, т.е. $F_{\xi}(x) = 0$.

2. $0 < x \leq 1$. $F_{\xi}(x) = P(\xi < x) = P(\xi = 0) = P(\Gamma\Gamma) = \frac{1}{4}$, тъй като събитието $\xi < x$ настъпва само, ако е настъпило събитието $\xi = 0$.
3. $1 < x \leq 2$. $F_{\xi}(x) = P(\xi < x) = P(\xi = 0) + P(\xi = 1) = P(\Gamma\Gamma) + P(\Gamma\mathcal{L}) = \frac{3}{4}$.
4. $x > 2$. Във всеки опит $\xi \leq 2 < x$, затова $F_{\xi}(x) = P(\xi < x) = 1$.

Графиката на функцията е начертана на фиг. 6.1. ♦



Фиг. 6.1

функцията на разпределение е непрекъсната.

§7. Дискретни случайни величини.

Множеството на възможните стойности на дискретните случайни величини е крайно или изброимо. Затова е удобно законът на разпределение на вероятностите да се задава не с функцията на разпределение $F_{\xi}(x)$, а чрез таблицата на разпределение,

$$\begin{array}{c|cccc} \xi & x_1 & x_2 & \dots & x_i & \dots \\ \hline p & p_1 & p_2 & \dots & p_i & \dots \end{array}$$

в която са описани всички възможни стойности $x_1, x_2, \dots, x_i, \dots$ на величината и вероятностите $p_i = P(\xi = x_i)$ в даден опит величината ξ да приеме стойността x_i . При това,

$$\sum p_i = 1, \quad (7.1)$$

където сумирането се извършва по цялото множество от възможни стойности на величината.

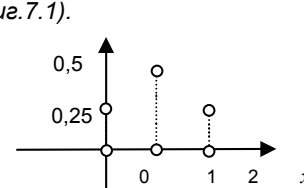
По този начин е определена функция

$$p_{\xi}(x_i) = P(\xi = x_i) = p_i, \quad i=1,2,\dots,$$

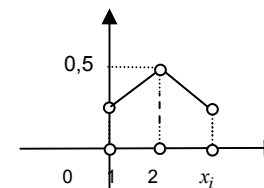
която ще наричаме плътност на вероятностите на дискретната случайна величина. Тя е дефинирана само за възможните стойности $x_1, x_2, \dots, x_i, \dots$ на величината ξ и има следните свойства:

$$1) \sum_i p_{\xi}(x_i) = 1; \quad 2) P(\alpha < \xi \leq \beta) = \sum_{x_i: \alpha < x_i \leq \beta} p_{\xi}(x_i),$$

а графиката ѝ е съвкупност от точки, с координати $(x_i, p_{\xi}(x_i))$ (фиг.7.1).



Фиг.7.1.



Фиг.7.2.

Освен с графиката на плътността на величината, законът на разпределение се изобразява графически и като начупена линия, свързваща точките с координати (x_i, p_i) , която се нарича многоъгълник на разпределението. На фиг. 7.1 и 7.2 са представени графиката на плътността и многоъгълникът на разпределението на случайната величина, разглеждана в пример 6.1.

Ако е дадена таблицата на разпределение, функцията $F_{\xi}(x)$ може да се получи като се използва, че събитието $\xi < x$ настъпва само, ако ξ е приела някоя от възможните си стойности, по-малки от x . Следователно, $F_{\xi}(x) = \sum_{x_i < x} p_i$ ($p_i = p_{\xi}(x_i)$). По този начин получаваме:

$$F_{\xi}(x) = \begin{cases} 0, & x \leq x_1 \\ p_1, & x_1 < x \leq x_2 \\ p_1 + p_2, & x_2 < x \leq x_3 \\ \dots & \dots \\ p_1 + \dots + p_{k-1}, & x_{k-1} < x \leq x_k \\ \dots & \dots \end{cases} \quad \text{като } \lim_{x \rightarrow \infty} F_{\xi}(x) = 1.$$

Функцията $F_{\xi}(x)$ е непрекъсната отдясно и за произволни a и b имаме

$$P(a \leq \xi < b) = F_{\xi}(b) - F_{\xi}(a).$$

$$P(\xi = a) = \lim_{b \rightarrow a, b > a} [F_{\xi}(b) - F_{\xi}(a)].$$

На фиг. 6.1 е дадена графика на функцията на разпределение на дискретна случайна величина от пример 6.1.

Ще разгледаме няколко примера, в които ще получим плътността и функцията на разпределение на някои дискретни случайни величини.

Нека A е случайно събитие и $P(A) = p$, $0 \leq p \leq 1$. Величината ξ със стойности 0 и 1, която показва колко пъти е настъпило събитието в един опит, се нарича индикатор на събитието A .

Пример 7.1. Да се състави таблицата на разпределение и да се намери функцията на разпределение на индикатора ξ на събитието A , за което $P(A)=p$.

Решение. Пресмятаме $P(\xi=1)=P(A)=p$, $P(\xi=0)=P(\bar{A})=1-p=q$, където q е вероятността за настъпване на противоположното събитие $\bar{A}$.

Следователно, $\frac{\xi}{P} \left| \begin{array}{c} 0 \\ q \end{array} \right| \frac{1}{p}$. Очевидно, $q+p=(1-p)+p=1$.

Тъй като възможните стойности на величината ξ са 0 и 1, то:

- За всяко число $x \in (-\infty, 0]$ имаме $F_{\xi}(x)=P(\xi < x)=0$ ($\xi < x$ е невъзможно).
- За число $x \in (0, 1]$ събитието $\{\xi < x\}$ настъпва само, ако е настъпило събитието $\{\xi=0\}$, следователно $F_{\xi}(x)=P(\xi < x)=P(\xi=0)=1-p$.
- За $x \in (1, \infty)$ имаме $F_{\xi}(x)=P(\xi < x)=P(\xi=0)+P(\xi=1)=q+p=1$.

По този начин, $F_{\xi}(x)=\begin{cases} 0, & x \in (-\infty, 0] \\ q, & x \in (0, 1] \\ 1, & x \in (1, \infty) \end{cases}$. ♦

Пример 7.2. Нека вероятността за поява на едно събитие A във всеки опит е $P(A)=p$. Извършваме опити докато се появи събитието A . Да се намери плътността на величината ξ - брой на направените опити.

Решение. Възможните стойности на величината са $1, 2, 3, \dots, n, \dots$ (изброимо много). Ако означим $P(\bar{A})=1-p=q$, то получаваме

$$p_{\xi}(1)=P(\xi=1)=P(A)=p, \quad p_{\xi}(2)=P(\xi=2)=P(\bar{A} \cap A)=qp,$$

$P(\xi=3)=P(\bar{A} \cap \bar{A} \cap A)=q^2p$ и т.н. Така получаваме, че плътността на величината е $p_{\xi}(n)=pq^{n-1}$, $n=1, 2, \dots$, която представяме с таблицата

$$\frac{\xi}{P} \left| \begin{array}{cccccc} 1 & 2 & 3 & \dots & n & \dots \\ p & pq & pq^2 & \dots & pq^{n-1} & \dots \end{array} \right. \quad (7.2)$$

Ще проверим равенство (7.1). Действително, имайки предвид, че $|q|<1$ и формулата за сума от безкрайна геометрична прогресия,

$$\text{получаваме} \quad p + pq + pq^2 + \dots + p^{n-1} + \dots = p \cdot \frac{1}{1-q} = p \cdot \frac{1}{p} = 1. \quad \blacklozenge$$

Тъй като вероятностите в таблица (7.2) са членове на геометрична прогресия, полученото разпределение се нарича геометрично. То зависи от един параметър p . Например броят на подхвърлянията на монета до поява на герб е геометрично разпределена величина с параметър $p=0,5$, а броят на хвърлянията на зар до поява на шестца – с параметър $p=1/6$.

Ако са дадени една или няколко случайни величини, то от тях може да образуваме нови случайни величини, например,

Сума $\zeta = \xi + \eta$ на две случайни величини ξ и η ще наричаме случайната величина ζ , която према стойност, равна на сумата от приетите стойности на величините ξ и η , т.е. $\zeta(\omega) = \xi(\omega) + \eta(\omega)$ за всяко $\omega \in \Omega$. Аналогично се дефинира произведение $\xi\eta$.

Могат да се дефинират и други неслучайни зависимости от случайни величини, например, $\psi = \sin \xi$, $\zeta = (\xi - E\xi)(\eta - E\eta)$.

Пример 7.3. Има две партиди изделия. Произволно изделие може да бъде дефектно с вероятност 0,1 за първата партида и 0,2 за втората. От всяка партида се избират по две изделия за проверка.

а) Да се намерят законите за разпределение на случайните величини:

ξ - брой на проверените дефектни изделия от първата партида

η - брой на проверените дефектни изделия от втората партида.

б) да се намери законът на разпределение на величината $\zeta = \xi + \eta$ - общ брой на проверените дефектни изделия.

в) За всяко дефектно изделие се плаща неустойка от 10 лв. Да се намери разпределението на величината σ - общата сума на изплатените неустойки.

Решение. Възможните стойности на величината ξ са 0, 1 или 2. За да изчислим вероятностите им, въвеждаме случайните събития $A_1 = \{\text{първото проверено изделие е дефектно}\}$ и $A_2 = \{\text{второто проверено изделие е дефектно}\}$. Тогава

$$P(\xi=0)=P(\bar{A}_1 \cap \bar{A}_2)=0,9 \cdot 0,9=0,81, \quad P(\xi=2)=P(A_1 \cap A_2)=0,1 \cdot 0,1=0,01,$$

$$P(\xi=1)=P((A_1 \cap \bar{A}_2) \cup (\bar{A}_1 \cap A_2))=0,1 \cdot 0,9 + 0,9 \cdot 0,1=0,09+0,09=0,18$$

Аналогични изчисления правим и за случайната величина η , с което получаваме таблиците

$$\frac{\xi}{P} \left| \begin{array}{ccc} 0 & 1 & 2 \\ 0,81 & 0,18 & 0,01 \end{array} \right|, \quad \frac{\eta}{P} \left| \begin{array}{ccc} 0 & 1 & 2 \\ 0,64 & 0,32 & 0,04 \end{array} \right|.$$

б) Възможните стойности на величината ζ са 0, 1, 2, 3 и 4. Като използваме формулите за сума и произведение на вероятностите, получаваме следната таблица на разпределение (проверете!)

$$\frac{\zeta}{P} \left| \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \\ 0,5184 & 0,3744 & 0,0964 & 0,0104 & 0,0004 \end{array} \right|.$$

в) Очевидно, възможните стойности на сумата, която трябва да се плати като неустойки, се изразява по формулата $\sigma=10\zeta$. Възможните й стойности са 0, 10, 20, 30, и 40 лв и имат същите вероятности, каквито са вероятностите на възможните стойности на величината ζ . ♦

Ако дадена случайна величина има възможна стойност, вероятността на която е най-голяма, то тази стойност се нарича мода.

Една случайна величина може да няма мода (например, равномерното разпределение от пример 8.2), а може да има повече от една мода.

Например, модата на величините ξ , η и ζ в пример 7.3. е възможната стойност 0.

Упражнения.

1. Двама души подхвърлят монета. Ако при дадено подхвърляне се падне "герб", то първият играч получава 1 лев, а ако се падне "лице", той дава 1 лев. Да се намерят законът и функцията на разпределение на величината ξ - печалбата на първия играч при едно хвърляне на монетата. Какво разпределение ще има величината ζ - печалба на първия играч при n хвърляния на монетата?

2. Зарче се хвърля до падане на 6 точки. Да се състави законът за разпределение на броя ξ на направените опити.

3. Зарче се хвърля 3 пъти. Да се състави законът за разпределение на ξ - брой на хвърляния, при които са се паднали не по-малко от 5 точки,

4. В партида от 10 изделия има 1 нестандартно. Изделията се проверяват докато се появи нестандартното. Да се състави законът на разпределение на броя ξ на проверените изделия, ако проверката се извършва: а) без връщане б) с връщане.

§8. Математическо очакване и други числени характеристики на дискретна случайна величина.

Законът за разпределение на дискретна величина, зададен с функцията на разпределение или таблицата на разпределение определя напълно една случайна величина. Но освен детайлно определяне на вероятността за приемане на всяка от възможните й стойности, в много случаи е важно да се определи каква е средната стойност на величината. Например, ако са дадени таблиците на разпределение на броя попадения на двама стрелци

ξ	0	1	2	3	4
P	0,1	0,1	0,2	0,1	0,5

η	0	1	2	3	4
P	0,1	0,1	0,2	0,2	0,4

е трудно да се определи кой е по-добрият, но ако си представим че всеки от тях е стрелял по 100 пъти, то може да очакваме, че средният брой попадения за всеки е

$$\frac{10.0+10.1+10.2+20.3+50.4}{100}=2,90 \text{ и } \frac{10.0+10.1+20.2+20.3+40.4}{100}=2,70.$$

Следователно, по-добър е първият стрелец.

Ето защо за изследването на общото поведение на случайните величини се въвеждат няколко числени характеристики, основна за които се явява характеристиката, наречена математическо очакване.

Нека дискретната случайна величина ξ е зададена с таблицата си на разпределение

ξ	x_1	x_2	...	x_n	...
P	$p_\xi(x_1)$	$p_\xi(x_2)$	...	$p_\xi(x_n)$	...

Математическо очакване на случайна величина ξ се нарича реално число $E\xi$ (други означения $E\xi$, $m\xi$, $M\xi$, m_ξ), определено по формулата

$$E\xi = \sum x_i p_\xi(x_i), \quad (8.1)$$

където сумирането се извършва по всички възможни стойности x_i .

Математическо очакване може и да не съществува, ако редът (8.1) е разходящ.

Доказано е експериментално, че математическото очакване е близко до средната стойност на величината, измерена в серия от опити.

Свойства на математическото очакване:

- 1) ако C е константа, то $EC = C$.
- 2) ако C е константа, то $E(C\xi) = CE\xi$.
- 3) за произволни случайни величини ξ_1 и ξ_2 $E(\xi_1 \pm \xi_2) = E\xi_1 \pm E\xi_2$.
- 4) ако случайните величини ξ_1 и ξ_2 са независими, то $E(\xi_1 \xi_2) = E\xi_1 E\xi_2$.
- 5) Математическото очакване на случайната величина $\eta = f(\xi)$, където f е неслучайна функция, се пресмята по формулата

$$E\eta = E(f(\xi)) = \sum f(x_i) p_\xi(x_i)$$

Центрирана случайна величина се нарича величина, математическото очакване на която е равно на нула. Например, величината $\bar{\xi} = \xi - E\xi$ е центрирана.

Дисперсия. Друга важна числена характеристика е дисперсията, която дава мярка за разсейването на случайната величина около математическото й очакване.

Дисперсия на случайната величина ξ се нарича числото (ако съществува)

$$D\xi = E(\xi - E\xi)^2 = \sum (x_i - E\xi)^2 p_\xi(x_i) \quad (8.2)$$

(т.е. математическото очакване на квадрата на центрираната величина $\bar{\xi} = \xi - E\xi$).

Свойства на дисперсията:

- 1) $D\xi = E\xi^2 - (E\xi)^2$, където $E\xi^2 = \sum x_i^2 p_i$
- 2) $D\xi \geq 0$ за произволна случайна величина
- 3) ако C е константа, то $D\xi = 0$

- 4) ако C е константа, то $D(C\xi) = C^2 D\xi$
 5) ако ξ_1 и ξ_2 са независими, то $D(\xi_1 \pm \xi_2) = D\xi_1 + D\xi_2$.

Забележка 8.1. Формула (8.3) е много удобна за пресмятането на дисперсията.

Средно квадратично отклонение.

Величината $\sigma_\xi = \sqrt{D\xi}$ се нарича средно квадратично отклонение на величината ξ .

Пример 8.1. Да се намери средно квадратичното отклонение на броя ξ на падналите точки при хвърляне на един зар (виж пример 8.1).

Решение. Имаме $E\xi = \frac{1}{6}(1+2+3+4+5+6) = \frac{7}{2}$.

$$E\xi^2 = \frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = \frac{91}{6},$$

$$D\xi = E\xi^2 - (E\xi)^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12} = 2,92,$$

Следователно, $\sigma_\xi = \sqrt{D\xi} = \sqrt{2,92} = 1,71$.

Пример 8.2. (Продължение на пример 7.3). Да се намерят математическите очаквания на величините ξ и η . Колко лева средно трябва да се предвидят за неустойки?

Решение. Изчисляваме:

$$E\xi = 0,081 + 1,018 + 2,001 = 0,2, \quad E\eta = 0,064 + 1,032 + 2,004 = 0,4.$$

Средната стойност на разходите за неустойки е близка до математическото очакване, следователно е равна на

$$E\sigma = E[10(\xi + \eta)] = 10(E\xi + E\eta) = 10 \cdot 0,6 = 6 \text{ лв} \quad \blacklozenge$$

Пример 8.3. Хвърлят се n правилни зара. Да се намери математическото очакване и средно квадратичното отклонение на сумата ξ от падналите точки.

Решение. Ако ξ_k е броят на падналите точки на k -тия зар, то

имаме
$$\begin{array}{c|cccccc} \xi_k & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline P & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{array}$$
 . Следователно, $E\xi_k = \frac{7}{2}$, $D\xi_k = \frac{35}{12} = 2,92$,

$$\sigma_{\xi_k} = \sqrt{D\xi_k} = 1,71.$$

Сумата от падналите точки представяме като $\xi = \xi_1 + \xi_2 + \dots + \xi_n$.

Величините ξ_k са независими, затова:

$$E\xi = E(\xi_1 + \xi_2 + \dots + \xi_n) = E\xi_1 + E\xi_2 + \dots + E\xi_n = \frac{7n}{2}$$

$$D\xi = D(\xi_1 + \xi_2 + \dots + \xi_n) = D\xi_1 + D\xi_2 + \dots + D\xi_n = 2,92n$$

$$\sigma_\xi = \sqrt{D\xi} = \sqrt{2,92n} = 1,71\sqrt{n} \quad \blacklozenge$$

Величина, която приема всяка една от своите стойности с еднаква вероятност, се нарича равномерно разпределена.

Следователно, величините ξ_i са равномерно разпределени. Лесно се вижда, че тяхната сума не е равномерно разпределена (виж пр. 4.2).

Обобщения на разгледаните числени характеристики $E\xi$ и $D\xi$ са:

$$\alpha_k = E(\xi^k) = \sum x_i^k p_\xi(x_i) - \text{начални моменти от ред } k \text{ (} k=0,1,\dots \text{)}$$

$$\mu_k = E(\xi - E\xi)^k = \sum (x_i - E\xi)^k p_\xi(x_i) - \text{централни моменти от ред } k \text{ (} k=2,3,\dots \text{)}.$$

$$\begin{aligned} \text{В частност,} \quad & \mu_0 = 1, \quad \mu_1 = E\xi, \quad \mu_2 = E\xi^2, \\ & \nu_0 = 1, \quad \nu_1 = 0, \quad \nu_2 = D\xi \\ & \nu_2 = D\xi = \mu_2 - \mu_1^2, \\ & \nu_3 = \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3 \\ & \nu_4 = \mu_4 - 4\mu_1\mu_3 + 6\mu_1^2\mu_2 - 3\mu_1^4 \end{aligned} \quad (8.4)$$

Ще отбележим още две важни характеристики, свързани с моментите:

■ асиметрия $a_\xi = \frac{\mu_3}{\sigma_\xi^3}$ - характеризира асиметричността

■ ексцес $\varepsilon_\xi = \frac{\mu_4}{\sigma_\xi^4} - 3$ - характеризира стръмнината на графиката на p_ξ .

Упражнения.

1. При игра на рулетка се залага на червено или черно. Половината от числата от 1 до 36 са червени, а другата половина – черни. Играч, заложил един лев, печели един лев, ако се падне цветът, на който е заложил и губи един лев, ако се падне противоположният цвят или някое от числата 0 и 00. Да се намери математическото очакване на печалбата.

2. Да се намерят $D\xi$ и σ_ξ на величината
$$\begin{array}{c|cccccc} \xi & 0 & 1 & 2 & 3 & 4 \\ \hline P & 0,1 & 0,2 & 0,4 & 0,2 & 0,1 \end{array}$$
.

3. Случайната величина ξ има таблица на разпределение
$$\begin{array}{c|ccc} \xi & 0 & 1 & 2 \\ \hline P & \frac{1}{2} & \frac{1}{3} & \frac{1}{6} \end{array}$$
. Да се

намерят законите на разпределение на величините $\xi-1$, ξ^2 и $(\xi-1)^2$, математическите очаквания и дисперсиите на четирите случайни величини.

4. Дисперсиите на независимите случайни величини ξ_1 и ξ_2 са $D\xi_1 = 11$ и $D\xi_2 = 4$.

Да се намери средно квадратичното отклонение на $\eta = 2\xi_1 - 5\xi_2$, $\zeta = \eta - \xi_2$.

5. Законът за разпределение на случайната величина ξ е $\frac{\xi}{P} \begin{matrix} -1 & 0 & 1 & 2 \\ \frac{1}{16} & \frac{1}{4} & \frac{1}{2} & \frac{3}{16} \end{matrix}$. Да се

намери $P(\xi > 0)$, модата, $E\xi$ и $\sigma\xi$. Да се построят графиките на плътността на и на функцията на разпределението на величината.

§9. Непрекъснатата случайна величина. Функция и плътност на разпределение. Мода, медиана, квантили и критични точки.

В тази глава ще разгледаме начините за определяне закона на разпределение на непрекъснатата случайна величина, т.е. правилото, по което да определяме вероятността на произволно случайно събитие, свързано с такъв тип величина. Очевидно, основни са събития от вида $\{\xi = a\}$ – величината ξ да приеме дадена стойност a и от вида $\{a \leq \xi < b\}$ – величината да приеме стойност в даден интервал (затворен, отворен или полуотворен), с помощта на които може да се представи всяко друго събитие. Както вече знаем (вж. §6), законът на случайна величина може да се определи чрез

функция на разпределение (интегрална функция на разпределението),

$$F_{\xi}(x) = P(\xi < x) \quad x \in (-\infty, \infty),$$

за която предполагаме, че е непрекъсната за всяко x .

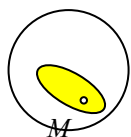
Първо ще покажем как може да се получи функцията на разпределение на една непрекъсната случайна величина.

Пример 9.1. Случайна частица M попада в кръг $k(O, R)$ с център O и радиус R . Вероятността частицата да попадне в произволна област D на кръга е пропорционална на лицето на областта. Измерва се разстоянието ξ до центъра O на кръга до частицата (фиг. 9.1). Да се намери и начертае графиката на функцията на разпределение на случайната величина ξ . Да се намерят вероятностите $P(-1 \leq \xi < 0)$, $P(R/3 \leq \xi < R/2)$.

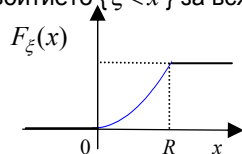
Решение. За вероятността на събитието $A = \{M \in D\}$ имаме

$P(A) = \frac{S(D)}{\pi R^2}$ (геометрична вероятност (§3)). Трябва да пресметнем

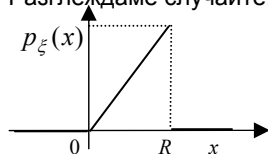
вероятността на събитието $\{\xi < x\}$ за всяко x . Разглеждаме случаите:



Фиг. 9.1.



Фиг.9.2.



Фиг.9.3.

1. $x \leq 0$. Тогава $F_{\xi}(x) = P(\xi < x) = 0$, тъй като във всеки опит ξ приема неотрицателна стойност.

2. При $0 < x \leq R$ събитието $\{0 < \xi \leq x\}$ означава, че точката M попада в

кръг с радиус x , т.е. $F_{\xi}(x) = P(M \in k(O, x)) = \frac{\pi x^2}{\pi R^2} = \frac{x^2}{R^2}$

3. При $R < x$ събитието $\{\xi < x\}$ винаги настъпва, защото R е най-голямата възможна стойност на ξ , откъдето $F_{\xi}(x) = 1$,

$$\text{Следователно, } F_{\xi}(x) = \begin{cases} 0, & x \leq 0 \\ \frac{x^2}{R^2} & 0 < x \leq R, \\ 1 & x > R \end{cases} \text{ графиката на която е дадена}$$

на фиг 9.2. Сега може да намерим търсените вероятности

$$P(-1 \leq \xi < 0) = F(0) - F(-1) = 0 - 0 = 0,$$

$$P\left(\frac{R}{3} \leq \xi < \frac{R}{2}\right) = F\left(\frac{R}{2}\right) - F\left(\frac{R}{3}\right) = \frac{1}{4} - \frac{1}{9} = \frac{5}{36} \cdot \blacklozenge$$

Функцията, която получихме в примера, има всички свойства, които би трябвало да има една функция на разпределение, а именно

- 1) $0 \leq F_{\xi}(x) \leq 1$ за всяко x ,
- 2) $F_{\xi}(x)$ е непрекъсната и не намаляваща за всяко x
- 3) $\lim_{x \rightarrow -\infty} F_{\xi}(x) = 0, \lim_{x \rightarrow \infty} F_{\xi}(x) = 1$,
- 4) $P(a \leq \xi < b) = F_{\xi}(b) - F_{\xi}(a)$

За да намерим вероятността на събитието $\{\xi = a\}$, съгласно забележка 6.1, извършваме граничния преход

$$P(\xi = a) = \lim_{b \rightarrow a} P(a \leq \xi < b) = \lim_{b \rightarrow a} [F_{\xi}(b) - F_{\xi}(a)] = 0.$$

Следователно, за непрекъснати случайни величини имаме:

$$P(\xi = a) = 0, \text{ за всяко } a$$

$$P(a < \xi < b) = P(a < \xi \leq b) = P(a \leq \xi \leq b) = P(a \leq \xi < b) = F_{\xi}(b) - F_{\xi}(a)$$

Както за дискретните случайни величини, така и за непрекъснатите, функцията на разпределение не винаги е удобна. За непрекъснатите случайни величини се определя и друга функция, наречена плътност:

Плътност на разпределението на непрекъснатата случайна величина ξ (диференциална функция на разпределението) наричаме функцията, определена от равенството

$$p_{\xi}(x) = F'_{\xi}(x).$$

От дефиницията следва, че плътността може да се представи и като

$$p_{\xi}(x) = \lim_{\Delta x \rightarrow 0} \frac{F_{\xi}(x + \Delta x) - F_{\xi}(x)}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{P(x \leq \xi < x + \Delta x)}{\Delta x}$$

Плътността на разпределение е функция, която е дефинирана за всяко x и която има свойствата:

$$1) p_{\xi}(x) \geq 0 \text{ за всяко } x; \quad 2) \int_{-\infty}^{\infty} p_{\xi}(x) dx = 1.$$

$$3) P(a \leq \xi < b) = \int_a^b p_{\xi}(x) dx \quad 4) F_{\xi}(x) = \int_{-\infty}^x p_{\xi}(t) dt$$

Следователно, между двете функции $p_{\xi}(x)$ и $F_{\xi}(x)$, характеризиращи вероятностно една непрекъсната случайна величина, съществуват следните връзки:

$$p_{\xi}(x) = F'_{\xi}(x), \quad F_{\xi}(x) = \int_{-\infty}^x p_{\xi}(t) dt.$$

Например, за пример 9.1. $p_{\xi}(x) = F'_{\xi}(x) = \begin{cases} 2x/R^2 & x \in [0, R) \\ 0 & x \notin [0, R) \end{cases}$.

Вероятността случайната величина ξ да попаде в интервала $[a, b]$ може да се пресметне по една от двете формули

$$P(a \leq \xi < b) = F_{\xi}(b) - F_{\xi}(a) \text{ или } P(a \leq \xi < b) = \int_a^b p_{\xi}(x) dx$$

Геометрична интерпретация на последната формула е, че вероятността $P(a \leq \xi < b)$ е равна на лицето на криволинейния трапец образуван от графиката на плътността в интервала $[a, b]$.

Пример. 9.2. $F_{\xi}(x) = a + b \arctg x$ е функцията на разпределение на случайната величина ξ (разпределение на Коши). Да се намерят константите a и b . За се намери плътността на вероятностите и да се построи графиката ѝ.

Решение. Съгласно свойствата на функцията на разпределението

$$\lim_{x \rightarrow -\infty} F_{\xi}(x) = 0 \Rightarrow a + b \lim_{x \rightarrow -\infty} \arctg x = 0 \Rightarrow a - b \frac{\pi}{2} = 0$$

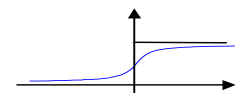
$$\lim_{x \rightarrow \infty} F_{\xi}(x) = 1 \Rightarrow a + b \lim_{x \rightarrow \infty} \arctg x = 1 \Rightarrow a + b \frac{\pi}{2} = 1,$$

откъдето получаваме $a = \frac{1}{2}$, $b = \frac{1}{\pi}$, следователно,

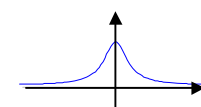
$$F_{\xi}(x) = \frac{1}{2} + \frac{1}{\pi} \arctg x.$$

За плътността на разпределение имаме $p_{\xi}(x) = F'_{\xi}(x) = \frac{1}{\pi(1+x^2)}$.

Графиките на двете функции са дадени на фиг.9.3 и фиг.9.4.



Фиг. 9.3



Фиг. 9.4

Плътността на разпределение дава по-ясна представа за характера на случайната величина, затова наименованието на закона е свързано с вида на функцията $p_{\xi}(x)$. Например, в пример 9.1 $p_{\xi}(x) > 0$ само в интервала $[0, R)$, което означава, че този интервал е множеството от възможните стойности на величината. В пример 9.2, тъй като $p_{\xi}(x) > 0$ за всяко x , теоретически величината може да приеме много големи или много малки стойности. От графиките на двете функции се вижда също, че величините не приемат всичките си стойности с еднаква вероятност, като за пример 9.1 най-голяма е тази вероятност за възможни стойности, близки до $x=R$, а за пример 9.2 $x=0$.

Възможната стойност M_o на величината ξ , за която $p_{\xi}(x) = \max$, се нарича мода.

Мода на величината в пример 9.2 е възможната ѝ стойност $x=0$.

Плътността на величина от пример 9.1 не достига максималната си стойност, затова тази величина няма мода.

В много практически задачи, особено тези, свързани със статистиката, при дадено число $0 < p < 1$ за величината ξ се търсят:

- възможната стойност x_p , за която $P(\xi < x_p) = p$. Тази стойност се нарича квантил от ред p и е решението на уравнението $F_{\xi}(x) = p$.
- възможната стойност $x_p^{(kp)}$, за която $P(\xi \geq x_p^{(kp)}) = p$. Нарича се критична точка от ред p . Очевидно, $x_p^{(kp)} = x_{1-p}$.

Най-често срещаните квантили носят свои названия:

Медиана M_e (Q_2) – квантил от ред 0,5, т.е. $M_e = Q_2 = x_{0,5}$.

Долен квантил Q_1 – квантил от ред 0,25.

Горен квантил Q_3 – квантил от ред 0,75.

Децили – квантили от редове 0,1, 0,2 и т.н.

Процентили – квантили от редове 0,01, 0,02 и т.н. (наричат се също 1%-процентил и т.н.)

Забележка 9.1. Въведените определения се отнасят и за дискретни случайни величини.

Пример 9.3. Да се намерят долният и горният квартил и критичната точка от ред 0,25 на величината ξ от пример 9.2.

Решение. Трябва да намерим квантилите $x_{0,25}$ и $x_{0,75}$, които са решенията на уравненията

$$F_{\xi}(x) = \frac{1}{4} \Rightarrow \frac{1}{2} + \frac{1}{\pi} \arctg x = \frac{1}{4} \Rightarrow \arctg x = -\frac{\pi}{4} \Rightarrow x = -\operatorname{tg} \frac{\pi}{4} = -1.$$

$$F_{\xi}(x) = \frac{3}{4} \Rightarrow \frac{1}{2} + \frac{1}{\pi} \arctg x = \frac{3}{4} \Rightarrow \arctg x = \frac{\pi}{4} \Rightarrow x = \operatorname{tg} \frac{\pi}{4} = 1.$$

Следователно, $x_{0,25} = -1$.

Тъй като $x_{0,25}^{(kp)} = x_{1-0,25} = x_{0,75}$, то $x_{0,75} = x_{0,25}^{(kp)} = 1$. ♦

Упражнения.

1. Нека опитът се състои в посочване на случайна точка M в интервала $[0,1]$ от оста Ox , при което се измерва разстоянието $OM = \xi$. Ако се предположи, че вероятността за попадане на точката в произволна отсечка е пропорционална на дължината ѝ, да се намерят функцията и плътността на разпределение на величината ξ .

2. Дадена е функцията на разпределение $F_{\xi}(x) = \begin{cases} 0 & x < -1 \\ a + b \arcsin x & -1 \leq x \leq 1 \\ 1 & x > 1 \end{cases}$. Да се намерят константите a и b . За се намери плътността на вероятностите и да се построи графиката ѝ.

3. Плътността на вероятностите на величината ξ е $p_{\xi}(x) = ae^{-|x|}$. Да се намери константата a и $P(-1 < \xi < 2)$.

4. Дадена е функцията на разпределение $F_{\xi}(x) = \begin{cases} 0, & x < 0 \\ 0,25x^2, & 0 \leq x \leq 2 \\ 1, & x > 2 \end{cases}$ на

случайната величина ξ . Да се намерят долният и горният квартил, медианата и критичните точки от ред 0,3 и 0,1.

5. Нека опитът се състои в случаен избор на точка P от отсечка AB с дължина единица като вероятността за попадение в произволен подинтервал е пропорционална на дължината на интервала. Измерва се разстоянието ξ от средата M на интервала до точка P .

а) Да се намери и начертае графиката на функцията на разпределение на случайната величина ξ .

б) Да се намерят законите на разпределение на величините ξ^{-1} , ξ^2 и $(\xi-1)^2$.

6. Плътността на разпределение на непрекъснатата случайна величина ξ е $p_{\xi}(x) = \begin{cases} kx^2(2-x) & x \in [0,2] \\ 0 & x \notin [0,2] \end{cases}$. Да се намери: а) стойността на k ; б) модата и медианата на величината ξ ; в) $P(1 \leq x \leq 2)$.

§10. Математическо очакване, дисперсия и други числени характеристики на непрекъснати случайни величини.

Нека непрекъснатата случайна величина ξ е зададена с плътността си $p_{\xi}(x)$ на разпределение. Математическото очакване, дисперсията и средно квадратичното отклонение се дефинират с формулите

$$E\xi = \int_{-\infty}^{\infty} xp_{\xi}(x)dx \quad - \text{математическо очакване} \quad (\text{други означения}$$

$$m\xi, M\xi, m_{\xi})$$

$$D\xi = E(\xi - E\xi)^2 = \int_{-\infty}^{\infty} (x - E\xi)^2 p_{\xi}(x)dx \quad - \text{дисперсия}$$

$$\sigma_{\xi} = \sqrt{D\xi} \quad - \text{средно квадратично отклонение.}$$

(ако съответните интеграли са сходящи).

Както при дискретните случайни величини, притежават свойствата:

На математическо очакване:

1. ако C е константа, то $EC = C$.
2. ако C е константа, то $E(C\xi) = CE\xi$.
3. за произволни величини ξ_1 и ξ_2 $E(\xi_1 \pm \xi_2) = E\xi_1 \pm E\xi_2$. (10.1)
4. ако случайните величини ξ_1 и ξ_2 са независими, то $E(\xi_1 \xi_2) = E\xi_1 E\xi_2$.
5. Математическото очакване на случайната величина $\eta = f(\xi)$, където f е неслучайна функция, се пресмята по формулата

$$E\eta = \int_{-\infty}^{\infty} f(x)p_{\xi}(x)dx$$

Дисперсия:

$$1. D\xi = E\xi^2 - (E\xi)^2, \text{ където } E\xi^2 = \int_{-\infty}^{\infty} p_{\xi}(x)x^2dx \quad (10.2)$$

2. $D\xi \geq 0$ за произволна случайна величина
3. ако C е константа, то $DC = 0$
4. ако C е константа, то $D(C\xi) = C^2 D\xi$
5. ако ξ_1 и ξ_2 са независими, то $D(\xi_1 \pm \xi_2) = D\xi_1 + D\xi_2$.

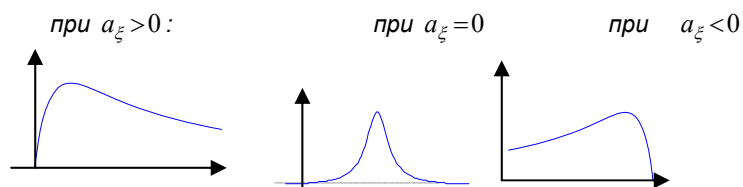
Обобщения на разгледаните числени характеристики $E\xi$ и $D\xi$ са:

- начални моменти от ред k : $\mu_k = \int_{-\infty}^{\infty} p_{\xi}(x)x^k dx \quad (k=1,2,\dots)$
- централни моменти от ред k : $\nu_k = \int_{-\infty}^{\infty} (x-E\xi)^k p_{\xi}(x)dx \quad (k=2,3,\dots)$

Централните моменти могат да се пресметнат с помощта на началните по формули (8.4)

Ще отбележим още две важни характеристики, свързани с моментите:

- асиметрия $a_{\xi} = \frac{\nu_3}{\sigma_{\xi}^3}$ - определя асиметричността на графиката:



ексцес $\varepsilon_{\xi} = \frac{\nu_4}{\sigma_{\xi}^4} - 3$ - характеризира стръмнината на графиката на $p_{\xi}(x)$

Пример 10.1. Да се намерят математическото очакване и дисперсията на величината с плътност $p_{\xi}(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & x \notin [a, b] \end{cases}$

(равномерно разпределена в интервала (a, b))

Решение. Пресмятаме:

$$E\xi = \int_{-\infty}^{\infty} xp_{\xi}(x)dx = \int_{-\infty}^a 0dx + \int_a^b \frac{1}{b-a}x dx + \int_b^{\infty} 0dx = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}.$$

$$E\xi^2 = \int_{-\infty}^{\infty} x^2 p_{\xi}(x)dx = \frac{1}{b-a} \int_a^b x^2 dx = \frac{(b^3 - a^3)}{3(b-a)} = \frac{b^2 + ab + a^2}{3}$$

$$D\xi = E\xi^2 - (E\xi)^2 = \frac{b^2 + ab + a^2}{3} - \frac{(b+a)^2}{4} = \frac{(b-a)^2}{12},$$

СЛУЧАЙНА ВЕЛИЧИНА – ОСНОВНИ РЕЗУЛТАТИ

1. Начини за определяне на случайните величини:

Дискретни случайни величини	Непрекъснати случайни величини
<u>функция на разпределение.</u> $F_{\xi}(x) = P(\xi < x)$ стъпаловидна ненамаляваща ф-ция. $0 \leq F_{\xi}(x) \leq 1$	<u>функция на разпределение.</u> $F_{\xi}(x) = P(\xi < x)$ ненамаляваща непр. ф-ция. $\lim_{x \rightarrow -\infty} F_{\xi}(x) = 0, \quad \lim_{x \rightarrow \infty} F_{\xi}(x) = 1$
<u>Плътност на разпределение</u> $p_{\xi}(x_i) = P(\xi = x_i) = p_i, \quad i=1,2,\dots$ $\begin{matrix} \xi \\ P \end{matrix} \left \begin{matrix} x_1 & x_2 & \dots & x_n \dots \\ p_{\xi}(x_1) & p_{\xi}(x_2) & \dots & p_{\xi}(x_n) \dots \end{matrix} \right.,$ $\sum p_{\xi}(x_i) = 1$ $F_{\xi}(x) = \sum_{x_i < x} p_{\xi}(x_i)$	<u>Плътност на разпределение</u> $\int_{-\infty}^{\infty} p_{\xi}(x)dx = 1, \quad p_{\xi}(x) \geq 0$ $p_{\xi}(x) = F'_{\xi}(x)$ $F_{\xi}(x) = \int_{-\infty}^x p_{\xi}(t)dt$
<u>Вероятност на събитието</u> $\{a \leq \xi < b\}$ $P(a \leq \xi < b) = F_{\xi}(b) - F_{\xi}(a) = \sum_{x_i: a \leq x_i < b} p_{\xi}(x_i)$	<u>Вероятност на събитието</u> $\{a \leq \xi < b\}$ $P(a \leq \xi < b) = F_{\xi}(b) - F_{\xi}(a) = \int_a^b p_{\xi}(x)dx$

2. Числени характеристики на случайните величини:

- математическо очакване: $E\xi = \begin{cases} \sum x_i p_{\xi}(x_i) & \xi - \text{дискр.} \\ \int_{-\infty}^{\infty} xp_{\xi}(x)dx & \xi - \text{непр.} \end{cases}$

Свойства на математическото очакване:

- 1) $EC = C$.
- 2) $E(C\xi) = CE\xi$
- 3) $E(\xi_1 \pm \xi_2) = E\xi_1 \pm E\xi_2$.
- 4) $E(\xi_1 \xi_2) = E\xi_1 E\xi_2$, ако ξ_1 и ξ_2 са независими.
- 6) Математическото очакване на случайната величина $\eta = f(\xi)$, където f е неслучайна функция, се пресмята по формулата

$$E\eta = \begin{cases} \sum f(x_k) p_{\xi}(x_k), & \xi - \text{дискретна.} \\ \int_{-\infty}^{\infty} f(x) p_{\xi}(x)dx, & \xi - \text{непрекъсната} \end{cases}$$

- дисперсия:** $D\xi = \begin{cases} \sum (x_i - E\xi)^2 p_\xi(x_i) & \xi - \text{дискретна} \\ \int_{-\infty}^{\infty} (x - E\xi)^2 p_\xi(x) dx & \xi - \text{непрекъсната} \end{cases}$

Свойства на дисперсията:

- $D\xi = E\xi^2 - (E\xi)^2$, 3) $D(C\xi) = C^2 D\xi$
- $DC = 0$ 4) $D(\xi_1 \pm \xi_2) = D\xi_1 + D\xi_2$, ако ξ_1 и ξ_2 са независими.

- Средно квадратично отклонение** $\sigma_\xi = \sqrt{D\xi}$

- начални моменти от ред k:** $\mu_k = \begin{cases} \sum x_i^k p_\xi(x_i) & \xi - \text{дискр.} \\ \int_{-\infty}^{\infty} x^k p_\xi(x) dx & \xi - \text{непр.} \end{cases}$

- централни моменти от ред k:** $\nu_k = \begin{cases} \sum (x_i - E\xi)^k p_\xi(x_i) & \xi - \text{дискр.} \\ \int_{-\infty}^{\infty} (x - E\xi)^k p_\xi(x) dx & \xi - \text{непр.} \end{cases}$

Връзки между централните и началните моменти:

$$\mu_0 = 1, \quad \mu_1 = E\xi, \quad \mu_2 = E\xi^2$$

$$\nu_2 = D\xi = \mu_2 - \mu_1^2$$

$$\nu_3 = \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3$$

$$\nu_4 = \mu_4 - 4\mu_1\mu_3 + 6\mu_1^2\mu_2 - 3\mu_1^4$$

- асиметрия:** $a_\xi = \frac{\nu_3}{\sigma_\xi^3}$

- ексцес:** $\varepsilon_\xi = \frac{\nu_4}{\sigma_\xi^4} - 3$

Общи задачи

1. Дадена е дискретната случайна величина ξ

ξ	0	1	2	3	4
P	p	$2p$	$4p$	$2p$	p

. Да се намерят: а) константата p ; б) $P(0 \leq \xi < 3)$; в) стойността на функцията на разпределение $F_\xi(x)$ за $x=3,4$, г) дисперсията.

2. $F_\xi(x) = \begin{cases} 0, & x < 0 \\ ax^2, & 0 \leq x \leq 2 \\ 1, & x > 2 \end{cases}$ е функцията на разпределение на величината ξ . Да се

намерят константата a , дисперсията, модата и медианата на ξ и вероятността величината ξ да приеме стойност, по-голяма от 1.

3. В една кутия има 4 бели и 2 черни топки. а) От кутията се изтеглят 3 топки по схемата с връщане. Намерете вероятностите на събитията $A=\{2 \text{ от топките са бели}\}$, $B=\{\text{поне 2 от тях са бели}\}$. б) От кутията се изтеглят 2 топки по схемата без връщане. За случайната величина ξ - брой на изтеглените бели топки да се намерят законът на разпределението, $E\xi$, $D\xi$ и σ_ξ^2 .

4. Да се намерят математическото очакване, дисперсията и средно квадратичното отклонение на случайните величини ξ_1 и ξ_2 с таблици на разпределение съответно

$$\xi_1 \begin{array}{c|ccccc} -2 & -1 & 0 & 1 & 2 \\ \hline P & 0,1 & 0,2 & 0,4 & 0,2 & 0,1 \end{array}, \quad \xi_2 \begin{array}{c|ccccc} -2 & -1 & 0 & 1 & 2 \\ \hline P & 0,05 & 0,3 & 0,3 & 0,3 & 0,05 \end{array}$$

5. Дадена е плътността $p_\xi(x) = a \cos x$ на величината ξ , която приема стойности в интервала $x \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. Да се намерят: а) константата a ; б) $P(|\xi| > \frac{\pi}{4})$; в) $E\xi$ г) функцията на разпределението на величината.

6. Плътността на разпределение на случайната величина ξ е

$$p_\xi(x) = \begin{cases} 0, & x < 3 \\ -\frac{3}{4}x^2 + \frac{9}{2}x - a, & 3 \leq x \leq 5 \\ 0, & x > 5 \end{cases}$$

Да се намерят: a , M_0 , M_e , $E\xi$, $P(3 < \xi < 4)$.
7. За един семестър студентите имат изпити по 4 дисциплини. Ако приемем, че броят ξ на изпитите, по които даден студент може да получи отличен е случайна величина с разпределение ξ

ξ	0	1	2	3	4
P	0,10	p	0,40	0,15	0,05

, да се намери: а)

вероятността студент да има само една отлична оценка; б) отличните оценки да са по-малко от 3; в) средният брой отлични оценки на един студент от курса; в) средно квадратичното отклонение на величината ξ .

8. Случайната величина ξ има таблица на разпределение ξ

ξ	0	1	2
P	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$

. Да се

намерят законите на разпределение на величините $\xi-1$, ξ^2 и $(\xi-1)^2$. Да се намерят математическите очаквания и дисперсиите на четирите случайни величини.

9. Дадена е плътността $p_\xi(x) = \begin{cases} k(1+x^2), & x \in [0,1] \\ 0, & x \notin [0,1] \end{cases}$ на непрекъснатата случайна

величина ξ . Да се изчислят: а) коефициентът k ; б) $P\left(0 \leq \xi \leq \frac{1}{3}\right)$; в) математическото очакване на величината; г) дисперсията $D\xi$.

10. Дадена е функцията $F_\xi(x) = \begin{cases} 0, & x < -1 \\ a(-x^3 + 3x + 2), & -1 \leq x \leq 1 \\ 1, & x > 1 \end{cases}$ на разпределение на

непрекъснатата случайна величина ξ . Да се намерят а) константата a ; б) $P(0 < \xi < 1)$; в) ексцесът и асиметрията на величината.

НЯКОИ ПО-ВАЖНИ ДИСКРЕТНИ И НЕПРЕКЪСНАТИ РАЗПРЕДЕЛЕНИЯ

§11. Биномно разпределение, разпределение на Поасон и геометрично разпределение.

Предстои да разгледаме някои разпределения на дискретни случайни величини, които се срещат често в практиката.

Ще припомним развитието на нютония бином

$$(a+b)^n = \sum_{k=0}^n \binom{n}{k} a^{n-k} b^k = a^n + \binom{n}{1} a^{n-1} b + \binom{n}{2} a^{n-2} b^2 + \dots + \binom{n}{k} a^{n-k} b^k + \dots + b^n$$

коэффициентите на който ще означаваме също и с C_n^k (виж §4) и са равни на

$$C_n^k = \binom{n}{k} = \frac{n(n-1)(n-2)\dots(n-k+1)}{k!}, \quad k=0,1,\dots,n, \quad C_n^0 = \binom{n}{0} = 1, \quad C_n^n = \binom{n}{n} = 1$$

Например,

$$C_n^1 = n, \quad C_n^2 = \frac{n(n-1)}{2}, \quad C_n^3 = \frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3}, \quad C_{14}^5 = \binom{14}{5} = \frac{14 \cdot 13 \cdot 12 \cdot 11 \cdot 10}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 2002.$$

Ще разглеждаме последователност от взаимно независими опити, т.е. такива опити, за които вероятността за настъпване на каквото и да е резултат във всеки от тях не зависи от това какви резултати са настъпили в останалите опити. Независимите опити могат да се провеждат в еднакви или в различни условия. В първия случай вероятността за появяване на някакво събитие е една и съща, а във втория се мени от опит в опит.

Нека опитите се извършват при еднакви условия и във всеки от опитите наблюдаваме появата на някакво събитие A . Очевидно, вероятностите $P(A)=p$ за поява (успешен опит) и $P(\bar{A})=1-p=q$ за непоява (неуспешен опит) на събитието A не се променят.

Изложената схема на повторения на опитите, при която във всеки опит вероятността $P(A)=p$ за събъждане на събитието A е постоянна, се нарича схема на Бернули.

Формула на Бернули. Ако вероятността p за събъждане на събитието A във всеки от n независими опита е постоянна, то вероятността за това, че в n опита събитието A ще настъпи k пъти е

$$C_n^k p^k q^{n-k}, \quad k=0,1,\dots,n,$$

където $q=1-p$ е вероятността на събитието $\bar{A}$.

Доказателство. Да означим с $1,2,\dots,n$ номерата на опитите. Ако събитието A е настъпило точно в k опита, то техните номера образуват една ненаредена k -торка (комбинация без повторения (§4)). Например основната комбинация $(1,2,\dots,k)$ означава, че събитието A е настъпило в

първите k опита, а в останалите $n-k$ опита - не е настъпило. Следователно, броят на различните случаи, в които A настъпва k пъти е C_n^k , като вероятността за всеки един от тези случаи е $p^k(1-p)^{n-k} = p^k q^{n-k}$ (опитите са независими), т.е.

$$P = \underbrace{p^k q^{n-k} + p^k q^{n-k} + \dots + p^k q^{n-k}}_{C_n^k \text{ пъти}} = C_n^k p^k q^{n-k} \cdot \diamond$$

Биномно разпределение.

Законът на разпределение на вероятностите на величина ξ , която приема възможни стойности $k=0,1,\dots,n$ с вероятности

$$P(\xi=k) = P_\xi(k) = C_n^k p^k q^{n-k}, \quad (q=1-p) \quad (11.1)$$

се нарича биномен с параметри n и p (бележи се също $\xi \sim B(n,p)$).

Математическото очакване и дисперсията на биомно разпределена величина ξ са

$$E\xi = np, \quad D\xi = npq. \quad (11.2)$$

Различните величини с биомно разпределение могат да бъдат разнообразни по смисъл, но може да бъдат обединени в случаите:

- ξ - брой на поява на A в n опита, ако вероятността $P(A)=p$ остава една и съща във всеки опит. Например, брой на попаденията в мишена от краен брой опита, брой на успешните опити при състезание.
- ξ - брой на обектите, притежаващи дадено свойство, ако обектите са краен брой и вероятността който и да е обект да притежава свойството е p . Например, брой на нестандартните изделия в дадена партида.

Получените в пример 5.3 разпределения са биомни с параметри $p=0,1$, $n=2$ и $p=0,2$, $n=2$.

Доказва се, че модата $M_o=k_0$ на величината ξ (за която вероятността $P_n(k)$ е най-голяма), се определя като цяло число, удовлетворяващо неравенството

$$np + p - 1 \leq k_0 \leq np + p. \quad (11.3)$$

Формулите (11.2) за математическото очакване и дисперсията на величината ξ (брой на поява на A в n опита) ще получим като представим величината по друг начин. Нека ξ_i е индикатор (пример 7.1) на събитието A за i -тия опит. Законът на разпределение на ξ_i е

$$\frac{\xi_i}{P} \left| \begin{array}{c} 0 \\ q \end{array} \right| \frac{1}{p}, \quad E\xi_i = p, \quad D\xi_i = q(0-p)^2 + p(1-p)^2 = qp(q+p) = qp.$$

Тъй като ξ_i приема единица, когато събитието A настъпи, то броят на поява на A в n опита е $\xi = \xi_1 + \xi_2 + \dots + \xi_n$.

Тогава, съгласно свойствата на математическото очакване и дисперсията имаме

$$E\xi = E\xi_1 + E\xi_2 + \dots + E\xi_n = np,$$

$$D\xi = D\xi_1 + D\xi_2 + \dots + D\xi_n = npq. \spadesuit$$

Пример 11.1. Стреля се 14 пъти по обект, вероятността за попадение в който при единичен изстрел е 0,2. Да се изчисли: а) най-вероятният брой попадения и неговата вероятност; б) вероятността за унищожаване на обекта, ако за това са необходими не по-малко от 4 попадения.

Решение. Да означим с ξ броя на попаденията. Тази величина има биномно разпределение с параметри $p=0,2$ и $n=14$, (т.е. $\xi \sim B(14, 0,2)$), следователно, $P(\xi=k) = P_\xi(k) = C_{14}^k 0,2^k 0,8^{14-k}$, $k=0,1,\dots,14$.

а) По формула (8.2) изчисляваме най-вероятния брой:

$14 \cdot 0,2 - 0,8 \leq k_0 \leq 14 \cdot 0,2 + 0,2 \Rightarrow 2 \leq k_0 \leq 3$, т.е. $\xi=2$ или $\xi=3$, за които

$$P_{14}(2) = C_{14}^2 0,2^2 0,8^{12} \approx 0,25, \quad P_{14}(3) = C_{14}^3 0,2^3 0,8^{11} \approx 0,25.$$

б) $P(\xi \geq 4) = 1 - P(\xi < 4) = 1 - P_{14}(0) - P_{14}(1) - P_{14}(2) - P_{14}(3) \approx 0,302. \spadesuit$

Забележка 11.1. Очевидно, ако събитието A се появява k пъти в серия от n опита, то събитето $\bar{A}$ се появява $n-k$ пъти, т.е. за биномно разпределените величини $\xi \sim B(n, p)$ - брой на поява на A и $\bar{\xi} \sim B(n, q)$ - брой на поява на $\bar{A}$ е изпълнено $P(\xi=k) = P(\bar{\xi}=n-k) = C_n^k p^k q^{n-k}$.

Разпределение на Поасон. При големи стойности на n и малки стойности на p коефициентът C_n^k става много голям, а произведението $p^k q^{n-k}$ - много малко, с което грешките от изчисленията нарастват. Но, ако от $E\xi=np$ изразим $p=E\xi/n$, то

$$P_\xi(k) = \binom{n}{k} p^k q^{n-k} = \frac{n(n-1)\dots[n-(k-1)]}{k!} \left(\frac{E\xi}{n}\right)^k \left(1 - \frac{E\xi}{n}\right)^{n-k} =$$

$$= \frac{(E\xi)^k}{k!} \frac{n}{n} \left(1 - \frac{1}{n}\right) \dots \left(1 - \frac{k-1}{n}\right) \left(1 - \frac{E\xi}{n}\right)^n \left(1 - \frac{E\xi}{n}\right)^{-k} \rightarrow \frac{(E\xi)^k}{k!} e^{-E\xi} \text{ при } n \rightarrow \infty.$$

Следователно,

$$P_\xi(k) = C_n^k p^k q^{n-k} \approx \frac{\lambda^k}{k!} e^{-\lambda}, \text{ където } \lambda = E\xi = np. \quad (11.4)$$

По този начин получаваме нова формула за пресмятане на вероятностите, която зависи от един параметър, равен на математическото очакване на величината. С това величината с разпределение $B(n, p)$ при голямо n и малко p се доближава друго разпределение, което се нарича разпределение на Поасон.

Казваме, че дискретната случайна величина ξ е разпределена по закона на Поасон с параметър $\lambda > 0$ ($\xi \sim Po(\lambda)$), когато приема изброимо много стойности $0, 1, 2, \dots, n, \dots$ и

$$P(\xi=k) = p_\xi(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k=0, 1, \dots, n, \dots \quad (11.5)$$

Ще проверим, че и тук $\sum_{k=0}^{\infty} P(\xi=k) = 1$. Действително,

$$\sum_{k=0}^{\infty} P(\xi=k) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1, \text{ тъй като } \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \text{ е развитието на } e^{\lambda} \text{ в ред.}$$

Математическото очакване и дисперсията на величина ξ със разпределение по закона (11.5) са $E\xi = \lambda, \quad D\xi = \lambda$.

Законът на Поасон се нарича също закон на редките явления, защото случайна величина, която се състои в поява на събитие A в n опита при големи стойности на параметъра n и малки стойности на $P(A) = p \ll 1$, съгласно формула (11.4), се счита за разпределена по закона на Поасон с параметър $\lambda = np$ (математическото очакване на съответната биномна величина).

Друг, много често срещан случай на поасоново разпределение е така нареченият прост (поасонов) поток от събития - последователност от събития, които настъпват в случаен момент на интервала t от времето и която има свойствата:

- стационарност - вероятността $P(k)$ за настъпването на k събития за време t да зависи само от k и дължината t на интервала от времето;
- отсъствие на последиствие - $P(k)$ не зависи от това колко събития са настъпили преди началото на разглеждания интервал от време;
- ординарност - появяване на две и повече събития за малък интервал от време е практически невъзможно.

Средният брой a на събития, които настъпват за единица време, се нарича интензивност на потока.

При изпълнение на тези условия, величината ξ - брой на събитията, настъпили за интервал от време с дължина t , има разпределение на вероятностите (11.5), където $\lambda = at$ е средният брой събития, настъпващи за време t . Такива величини са, например, броят на обаждания по телефона, брой на колите, минаващи за време t на дадено кръстовище и др.

Пример 11.2. Вероятността за попадение в целта при всеки изстрел е равна на 0,001. Да се намери вероятността от 1000 изстрела да има не по-малко от 2 попадения в целта (събитие A).

Решение. Ще приложим формулата на Поасон, за което изчисляваме $np = 1000 \cdot 0,001 = 1$, $e^{-\lambda} = e^{-1} = 0,3679$, т.е. $P_{1000}(k) \approx \frac{0,3679}{k!}$, $k = 0, 1, 2, \dots$ Тогава $P(\xi \geq 2) = 1 - P(\xi < 2) = 1 - P(\xi = 0) - P(\xi = 1) = 0,2641$.

Пример 11.3. Средният брой поръчки за такси, постъпващи в диспечерски пункт, е равен на 3. Да се намери вероятността, че за 2 минути ще постъпят: а) 4 ; б) по-малко от 4; в) поне една поръчка. ($e^{-6} = 0,0025$). Отг. а) 0,1339; б) 0,1522; в) 0,8478.

Накрая ще отбележим и други разпределения на дискретни случайни величини, някои от които вече разглеждахме в примери.

Геометрично разпределение.

Геометрично разпределение с параметър $0 < p < 1$ - величина с възможни стойности $1, 2, \dots, n, \dots$ и вероятности $P(\xi = i) = pq^{i-1}$, където

$$q = 1 - p. \text{ За тази величина } E\xi = \frac{1}{p}, D\xi = \frac{q}{p^2}.$$

Такава е величината ξ - брой на проведените опити до поява на събитието A , вероятността на което във всеки опит е $P(A) = p$, разгледана в пример 7.2.

Геометричното разпределение играе важна роля в теорията на масовото обслужване.

Например, група от хора чакат пред дадено гише. Често се приема, че вероятността да пристигне още един човек за единица време е p , а да не пристигне е $q = 1 - p$. Тогава, времето T на пристигане на следващият клиент е величина с геометрично разпределение като събитието $T = k$ означава, че следващият клиент е пристигнал след k минути.

Хипергеометрично разпределение. (виж пример 4.2.)

Хипергеометрично разпределение с параметри натуралните числа $N, M \leq N, n \leq N$ — величина ξ с възможни стойности

$$i = 0, 1, 2, \dots, \min(M, n), \text{ за която } P(\xi = i) = \frac{C_M^i C_{N-M}^{n-i}}{C_N^n} \quad i = 0, 1, 2, \dots, \min(M, n).$$

$$\text{Тук } E\xi = np, D\xi = \frac{N-n}{N-1} np(1-p), \text{ където } p = \frac{M}{N}.$$

Типичен пример за величина с хипергеометрично разпределение е броят на познатите числа при игрите на спорт-тото 5 от 35, 6 от 49. Например, ако означим с ξ броят на познатите числа от играта "6 от 49", то $N = 49$ - общ брой на числата, $M = 6$ - брой на печелившите числа, $n = 6$ - брой на попълнените числа във фиша. Тогава възможните

стойности на ξ са $i = 0, 1, \dots, 6$ с вероятности $P(\xi = i) = \frac{C_6^i C_{43}^{6-i}}{C_{49}^6}$ (виж също

пример 4.3)

Упражнения.

- Вероятността за попадение в целта при един изстрел е p (събитие A). Да се намери законът за разпределение величината ξ - брой на попаденията в 4 опита.
- Какво е по-вероятно за един от двама равностойни шахматисти – да спечели : а) 3 от 4 или 5 от 8 партии; б) не по-малко от 3 от 4 партии или не по-малко 5 от 8 партии.
- Хвърлени са 6 правилни зарчета. Каква е вероятността за падането на: а) поне една; б) точно една; в) точно две единици.
- При предаването на съобщение вероятността за изкривяване на всеки знак е 0,1. Каква е вероятността съобщение от а) 5 знака; б) 300 знака да бъде прието правилно.
- Транспортират се 500 изделия като вероятността за повреда на изделие по пътя е 0,002. Да се намери вероятността, че са повредени: а) 3 изделия; б) по-малко от 3; в) повече от три; г.) поне едно изделие.
- Дадена е случайната величина $\xi \sim B(8, 1/3)$. Да се намери а) $P(\xi \leq 2)$, б) $P(\xi \geq 2)$.
- Потокът от заявки, постъпващи в телефонна централа, е прост поасонов поток с интензитет 30 обаждания на 1 час. Да се намери вероятността, че за 1 минута ще постъпят не по-малко от 2 заявки.
- Зарче се хвърля 5 пъти. Да се намери вероятността на събитията: а) нито веднаж не се пада шестца; б) поне веднъж се пада шестца; в) шестца се пада поне 4 пъти.
- Вероятността за печалба от един лотариен билет е 0,01. Да се определи колко билета трябва да се закупят, че с вероятност не по-малка от 0,95, да има поне един печеливш билет.
- За биномно разпределена величина е известно, че $E\xi = 3,75$, $\sigma\xi = 0,25\sqrt{15}$. Да се намери плътността на вероятностите и начертае графиката ѝ. Да се намери $P(2 \leq \xi < 4)$.
- Известно е, че вероятността за повреда на поне един елемент в устройство, състоящо се от 25 еднотипни елемента е 0,1. Да се намери вероятността за повреда на един елемент.
- В една кутия има 6 бели и 4 черни топки. а) От кутията се изтеглят 3 топки по схемата без връщане. Намерете вероятностите на събитията $A = \{\text{три от топките са бели}\}$, $B = \{\text{поне една от тях е бяла}\}$. б) От кутията се изтеглят 3 топки по схемата с връщане. За случайната величина ξ - брой на изтеглените бели топки да се намерят законът на разпределението, $E\xi$, $D\xi$ и $\sigma\xi$.
- В една кутия има 6 бели и 4 черни топки. От кутията се изтегля една топка. Ако изтеглената топка е бяла, то тя се връща в кутията, ако е черна се отстранява. а) Да се съставят законите на величините ξ_i - брой на тегленията до поява на i -тата черна топка и да се намерят математическите им очаквания ($i = 1, 2, 3, 4$). б) Да се намери математическото очакване на $\xi = \xi_1 + \xi_2 + \xi_3 + \xi_4$. Да се опише какъв е смисълът на величината ξ и нейното математическо очакване.
- Урна съдържа 3 червени и 7 черни топки. Ако се изтегли червена топка, тя се отстранява, а ако се изтегли черна – се връща. Средно колко пъти трябва да се теглят топки, докато се изтеглят и трите червени топки?
- Даден е законът на разпределение на дискретната величина ξ .

$$\begin{array}{c|cccc} \xi & -1 & 0 & 2 & 4 & 5 \\ \hline P & p & 0,3 & p & 0,2 & 0,3 \end{array}$$

Да се намерят: а) константата p ; б) $E\xi$, $D\xi$ и σ_ξ^2 , в) $P(0 \leq \xi < 3)$.

§12. Равномерно и показателно разпределение. Сума на две и повече случайни величини.

Ще разгледаме случайна величина ξ с плътност (виж пример 10.1)

$$p_\xi(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & x \notin [a, b] \end{cases} \text{ - равномерно разпределение в интервала } [a, b],$$

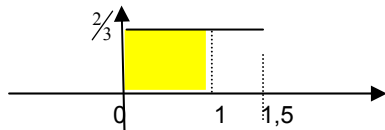
$$E\xi = \frac{a+b}{2}, \quad D\xi = \frac{(b-a)^2}{12}.$$

Равномерно разпределение имат величини, за които нямаме основание да смятаме, че приемат някои възможни стойности по-често от други.

Пример 12.1. На кръстовище е монтиран светофар, който свети 1 мин. зелено и 0,5 мин. червено. Автомобил се приближава към кръстовището в случаен момент, който не е свързан с работата на светофара. Да се намери вероятността, че автомобилът ще пресече кръстовището без да спира (събитие А).

Решение. Моментът ξ на пристигане на автомобилът е случайна величина, равномерно разпределена в интервала $(0; 1,5]$, равен на периода на смяна на светлината на светофара (фиг. 12.1).

Следователно, $P(A) = P(\xi < 1) = \frac{2}{3}$. ♦



Фиг.12.1.

2. Показателно (експоненциално разпределение) – величина ξ с плътност

$$p_\xi(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases} \text{ и } E\xi = \frac{1}{\lambda} \text{ и } D\xi = \frac{1}{\lambda^2}. \quad (12.2)$$

Показателно разпределение имат случайните величини:

- продължителност на работа на части от оборудване,
 - време на разпад на радиоактивен атом,
 - продължителност на телефонен разговор,
- най-общо казано, *времето до настъпване на някакво събитие.*

Пример 12.2. (Връзка между разпределението на Поасон и показателното разпределение). Да разгледаме прост поасонов поток от събития с интензивност λ (например пристигане на пътници на автобусна спирка) и величината T – времето от момента на наблюдение до първото настъпване на събитие (например времето от заминаването на поредния автобус до пристигането на първия пътник на спирката). Да се докаже, че величината T има показателно разпределение с параметър λ .

Решение. Интензивността λ на потока е равна на средния брой на настъпващи събития за единица време. Следователно, вероятността за време t да настъпят k събития (да пристигнат k пътника) е

$$P(k, t) = \frac{(\lambda t)^k e^{-\lambda t}}{k!} \text{ (Поасон)}.$$

Ще намерим функцията на разпределение $F_T(t) = P(T < t)$. За целта трябва да определим вероятността $P(T < t)$ за всяко t .

Ако $t < 0$, то, очевидно, $F_T(t) = 0$.

Ако $t \geq 0$, то имаме $P(T < t) = 1 - P(T \geq t)$. Събитието $P(T \geq t)$ означава, че за време t не е настъпило нито едно събитие, т.е. $P(T \geq t) = P(0, t) = e^{-\lambda t}$, следователно $F_T(t) = 1 - e^{-\lambda t}$.

От тук, $p_T(t) = F_T'(t) = \lambda e^{-\lambda t}$, $t \geq 0$ и $p_T(t) = 0$ при $t < 0$. ♦

Пример 12.3. Изследвайки броя на падналите метеорити в Сахара, е установено, че средно на 10 дни някъде в Сахара пада малък метеорит. Каква е вероятността, че метеорит ще падне в първия ден от наблюдението между 6 и 18 часа.

Решение. Случайната величина T – времето от започването на наблюдението до първото падане на метеорит има показателно разпределение, за което $ET = \frac{1}{\lambda} = 10 \Rightarrow \lambda = 0,1 \Rightarrow p_T(t) = 0,1e^{-0,1t}$, $t > 0$,

$$F_T(t) = 1 - e^{-0,1t}. \text{ Тогава } P\left(\frac{1}{4} < T < \frac{3}{4}\right) = F_T\left(\frac{3}{4}\right) - F_T\left(\frac{1}{4}\right) = 0,0479. \text{ ♦}$$

Сума на две и повече независими непрекъснати случайни величини. Да разгледаме сумата $\zeta = \xi + \eta$ на две случайни величини. Величината ζ приема възможна стойност $\zeta = z$, ако сумата от приетите стойности от величините ξ и η е равна на z . В такъв случай, ако величината ξ е приела стойност $\xi = u$, то величината η трябва да приеме стойността $\eta = z - u$. Така се получава, че плътността на величина ζ е

$$p_\zeta(z) = \int_{-\infty}^{\infty} p_\xi(u) p_\eta(z-u) du,$$

където $p_\xi(x)$ и $p_\eta(x)$ са плътностите на величините ξ и η (функцията

$p_{\zeta}(x)$ се нарича конволюция на функциите $p_{\xi}(x)$ и $p_{\eta}(x)$).

Както вече знаем, ако случайните величини ξ и η са независими, за математическото очакване и дисперсията на величината $\zeta = \xi + \eta$ имаме

$$E\zeta = E\xi + E\eta, \quad D\zeta = D\xi + D\eta.$$

Пример 12.4. Като използваме тези резултати, ще намерим плътностите на величините, които са сума на две, три и повече независими величини, имащи показателно разпределение с един и същ параметър λ .

Решение. Нека величините $\xi_1, \xi_2, \dots, \xi_n, \dots$ имат плътност на разпределение (12.2).

1) За плътността на $\zeta_2 = \xi_1 + \xi_2$ имаме $p_{\zeta_2}(z) = \int_{-\infty}^{\infty} p_{\xi_1}(u) p_{\xi_2}(z-u) du$,

където z е произволно. Подинтегралната функция е различна от нула, ако u е решение на системата неравенства $\begin{cases} u \geq 0 \\ z-u \geq 0 \end{cases}$, т.е. $0 \leq u \leq z$, затова

$$p_{\zeta_2}(z) = \int_{-\infty}^{\infty} p_{\xi_1}(u) p_{\xi_2}(z-u) du = \int_0^z \lambda e^{-\lambda u} \lambda e^{-\lambda(z-u)} du = \lambda^2 e^{-\lambda z} \int_0^z du = \lambda^2 e^{-\lambda z} z.$$

$$\text{Следователно, } p_{\zeta_2}(z) = \begin{cases} \lambda^2 z e^{-\lambda z} & z \geq 0 \\ 0 & z < 0. \end{cases}$$

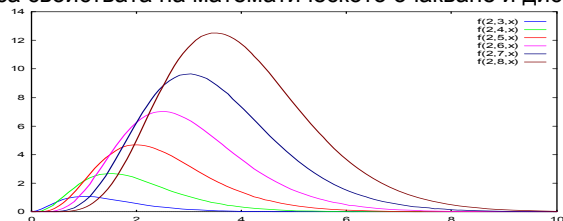
2) За плътността на $\zeta_3 = \xi_1 + \xi_2 + \xi_3 = \zeta_2 + \xi_3$ по аналогичен начин получваме

$$p_{\zeta_3}(z) = \int_{-\infty}^{\infty} p_{\zeta_2}(u) p_{\xi_3}(z-u) du = \int_0^z \lambda^2 e^{-\lambda u} \lambda e^{-\lambda(z-u)} u du = \frac{\lambda^3 z^2}{2!} e^{-\lambda z} \text{ за } z \geq 0.$$

По индукция може да се докаже, че величината $\zeta_n = \xi_1 + \xi_2 + \dots + \xi_n$ има плътност на разпределение

$$p_{\zeta_n}(x) = \begin{cases} \frac{\lambda^n x^{n-1}}{(n-1)!} e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}, \quad E\zeta_n = \frac{n}{\lambda}, \quad D\zeta_n = \frac{n}{\lambda^2},$$

(приложени са свойствата на математическото очакване и дисперсията).



Фиг. 12.2

Графиките на плътностите на разпределение при $\lambda=2$ и $n=3, \dots, 7$ са дадени на фиг 12.2

§13. Нормален закон на разпределение.

Ще разгледаме една величина, която има изключително значение както за теорията на вероятностите, така и за статистиката.

За величината ξ с плътност на разпределение

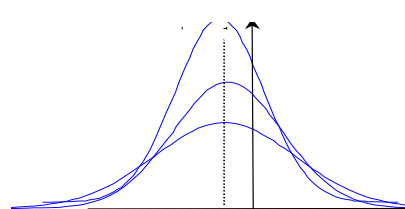
$$p_{\xi}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}} \quad (13.1)$$

се казва, че е подчинена на нормален закон (закон на Гаус) с параметри a и σ . Параметрите a и σ съвпадат с основните числени характеристики

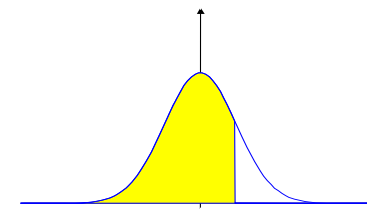
$$a = E\xi, \quad \sigma = \sigma_{\xi} = \sqrt{D\xi}.$$

$$a_{\xi} = 0, \quad e_{\xi} = 0$$

За краткост казваме, че величината с плътност (13.1) има $N(a, \sigma)$ - разпределение ($\xi \sim N(a, \sigma)$). От графиката на $p_{\xi}(x)$, дадена на фиг. 13.1, се вижда, че максимумът ѝ съвпада с математическото очакване, а кривата е толкова по-стръмна, колкото по малко е средно квадратичното отклонение. Абсциси на инфлексните точки на функцията са числата $a - \sigma$ и $a + \sigma$.



Фиг. 13.1.



Фиг. 13.2.

Величината със закон на разпределение $N(0,1)$, се нарича стандартна нормална величина и обикновено се означава с Z . Графиката на плътността

$$\varphi(x) = p_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

на стандартната нормална величина е дадена на фиг. 13.2. Лицето на заштрихованата област е равно на стойността на функцията на разпределение

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt, \quad F(-x) = 1 - F(x).$$

Стойностите на функцията $F(x)$ за $x \geq 0$ се вземат от таблица (стр.168).

От свойствата на функцията на разпределение следват равенствата:

$$P(Z < \beta) = F(\beta), \quad P(Z > \alpha) = 1 - F(\alpha),$$

$$P(\alpha < Z < \beta) = F(\beta) - F(\alpha)$$

Например, $P(Z < 1,46) = F(1,46) = 0,9279$;

$$P(Z \geq 2,21) = 1 - F(2,21) = 1 - 0,9864 = 0,0136$$
;

$$P(Z < -0,33) = F(-0,33) = 1 - F(0,33) = 1 - 0,6179 = 0,3821$$
;

$$P(0 < Z < 1) = F(1) - F(0) = 0,8413 - 0,5 = 0,3413.$$

За нормално разпределената величина $\xi \sim N(a, \sigma)$ е в сила формулата:

$$P(\alpha < \xi < \beta) = F\left(\frac{\beta - a}{\sigma}\right) - F\left(\frac{\alpha - a}{\sigma}\right) \quad (13.2)$$

приложение на която са частните случаи:

$$P(\xi < \beta) = F\left(\frac{\beta - a}{\sigma}\right), \quad P(\alpha < \xi) = 1 - F\left(\frac{\alpha - a}{\sigma}\right)$$

$$P(|\xi - a| < \delta) = 2F\left(\frac{\delta}{\sigma}\right) - 1 \quad (13.3)$$

Последната формула дава вероятността величината ξ да се отклони от математическото си очакване не повече от δ . Например,

$$P(|\xi - a| < 2\sigma) = 2F\left(\frac{2\sigma}{\sigma}\right) - 1 = 2.F(2) - 1 = 0,9544.$$

$$P(|\xi - a| < 3\sigma) = 2F\left(\frac{3\sigma}{\sigma}\right) - 1 = 2.F(3) - 1 = 0,9973.$$

Тъй като 0,9973 е близко до 1, то в сила е така наричаното

Правило на трите сигми: за нормално разпределена величина е изпълнено приближеното равенство

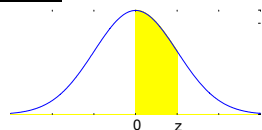
$$P(|\xi - a| < 3\sigma) \approx 1,$$

т.е. неравенството $|\xi - a| < 3\sigma$ е практически достоверно.

Забележка 13.2. В някои справочници е дадена таблица за

функцията на Лаплас

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-\frac{x^2}{2}} dx, \quad \Phi(-z) = \Phi(z)$$



За $x > 0$ между двете функции имаме връзката

$$F(x) = 0,5 + \Phi(x)$$

Функцията на Лаплас се използва по същия начин - за величина със закон на разпределение $N(a, \sigma)$ имаме

$$P(\alpha < \xi < \beta) = \Phi\left(\frac{\beta - a}{\sigma}\right) - \Phi\left(\frac{\alpha - a}{\sigma}\right), \quad P(|\xi - a| < \delta) = 2\Phi\left(\frac{\delta}{\sigma}\right)$$

Фундаменталната роля, която има нормалното разпределение се обяснява с това, че сума от независими случайни величини с нарастването на броя на събираемите асимптотически се доближава до нормално разпределена величина, което се доказва в централната гранична теорема (§15).

Друго свойство на нормалното разпределение:

Теорема 13.1. Ако са дадени две величини с нормално разпределение $X_1 \in N(a_1, \sigma_1)$ и $X_2 \in N(a_2, \sigma_2)$, то сумата им е нормално разпределена величина с математическо очакване $a_1 + a_2$ и дисперсия $\sigma_1^2 + \sigma_2^2$, т.е.

$$X_1 + X_2 \in N(a_1 + a_2, \sqrt{\sigma_1^2 + \sigma_2^2}).$$

Теоремата може да се обобщи за повече събираеми.

Пример 13.1. Големината ξ на произвежданите от завод детайли е нормално разпределена случайна величина с $a = 1,5 \text{ cm}$ (проектна големина), $\sigma = 0,04 \text{ cm}$ (точност на изработката). а) Да се пресметне вероятността произведено изделие да е нестандартно, ако се допуска отклонение от проектната големина $\pm 0,07$. б) Какво отклонение от проектната дължина може да се очаква с вероятност 0,97?

Решение. а) $P(A) = P(|\xi - 1,5| > 0,07) = 1 - P(|\xi - 1,5| \leq 0,07)$.

$$\text{По формула (13.3)} \quad P(|\xi - 1,5| \leq 0,07) = 2F\left(\frac{0,07}{0,04}\right) - 1 = 2F(1,75) - 1.$$

От таблицата $F(1,75) = 0,9599$. Следователно, $P(A) = 1 - 0,96 = 0,04$.

б) Търсим за каква стойност на ε е изпълнено $P(|\xi - 1,5| < \varepsilon) = 0,97$.

Отново по формула (13.3):

$$P(|\xi - 1,5| < \varepsilon) = 2F\left(\frac{\varepsilon}{0,04}\right) - 1 = 0,97 \Rightarrow F\left(\frac{\varepsilon}{0,04}\right) = 0,985,$$

откъдето от таблицата определяме $\frac{\varepsilon}{0,04} = 2,17 \Rightarrow \varepsilon \approx 0,09$.

Пример 13.2. Случайната величина ξ е подчинена на закона $N(2,1)$. Да се намери плътността на вероятностите и вероятността на неравенството $0 < \xi < 3$.

Упражнения.

1. Да се изчислят: $F(1,2)$, $F(-2)$, $F(2,34)$, $F(-0,75)$.
2. За величината $Z \sim N(0,1)$ да се изчислят вероятностите:
 $P(1,5 < Z < 2,34)$, $P(-1,2 < Z < 2,2)$, $P(-2,5 < Z < 2,5)$, $P(Z < 1,44)$,
 $P(Z < -2,2)$, $P(Z > 0,35)$, $P(Z > -2,6)$.
3. Дадена е случайната величина $X \sim N(36,10)$. Да се изчислят: а) $P(X > 48)$;
 $P(30 < X < 40)$; б) $P(|X - 36| < 0,03)$.
4. Дадена е величината $\xi \sim N(56,10)$. Да се намерят: $P(\xi > 68)$, $P(56 < \xi < 65)$,
 $P(42 < \xi < 52)$, $P(\xi < 36)$.
5. За нормално разпределената величина е дадено, че $P(\xi > 58,39) = 0,0217$,
 $P(\xi < 41,82) = 0,0287$. Да се намерят математическото очакване и средно
квадратичното отклонение на величината.
6. Да се намери x , ако а) $F(x) = 0,9660$, б) $F(x) = 0,3783$, в) $F(x) = 0,9495$,
г) $F(x) = 0,0224$, д) $F(x - 3) = 0,0591$, е) $F\left(\frac{x}{3}\right) = 0,8087$, ж) $F(2 - x) = 0,6217$.
7. Дължината на детайл е нормално разпределена величина с математическо
очакване 450 мм и средно квадратично отклонение 3 мм. Да се намери
вероятността произволно избран детайл да има дължина между 444 мм и 452 мм.
8. Установено е, че средната височина X на 11-годишните момчета е 146 см със
средно квадратично отклонение 8 см.
а) Какво разпределение може да да приемем за величината X ? Какъв е процентът
на момчетата, които са високи между 138 см и 154 см?
б) Какъв е процентът на момчетата, които са по-ниски от 130 см?
в) Какъв е процентът на момчетата, които са по-високи от 162 см?
9. Спортист счита, че времето, за което пробягва 200м, е нормално разпределена
случайна величина X като средното му постижение е 22,8 сек.
а) Ако е известно, че в 20% от случаите времето му е над 23,3 сек, да се
пресметне σX ;
б) Най-доброто постижение в клуба е 21,82. Да се намери вероятността
спортистът да надмине рекорда, ако $\sigma X = 0,42$

§14. Други разпределения на непрекъснати случайни величини

В тази глава ще използваме функция, наречена

Гама-функция: $\Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x} dx$ със свойства

$$1) \Gamma(1) = 1, \quad 2) \Gamma(z+1) = z\Gamma(z) \Rightarrow \Gamma(n+1) = n!, \quad n\text{-цяло}.$$

$$3) \Gamma(z)\Gamma(1-z) = \frac{\pi}{\sin \pi z}.$$

От свойство 3) се получава важното следствие

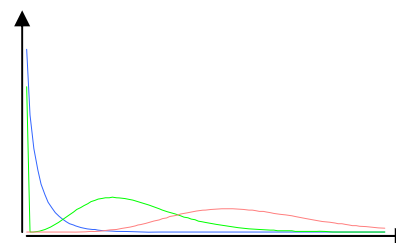
$$\int_0^{\infty} e^{-\frac{x^2}{2}} dx = \sqrt{\frac{\pi}{2}} \quad (\text{интеграл на Поасон}) \quad \left(\text{полагаме } z = \frac{1}{2}\right)$$

Ще разгледаме някои разпределения, които намират широко приложение в статистиката и които са свързани със стандартното нормално разпределение $N(0,1)$.

Хи-квадрат разпределение.

Ако $\xi_1, \dots, \xi_k \in N(0,1)$, то казваме, че величината $\xi = \xi_1^2 + \dots + \xi_k^2$ има χ^2 -разпределение с k степени на свобода. (хи-квадрат-разпределение) с означение $\xi \sim \chi^2(k)$.

Математическото очакване и дисперсията на величината са
 $E\xi = k$, $D\xi = 2k$.



Фиг.15.1.

Плътността на разпределение на величината е

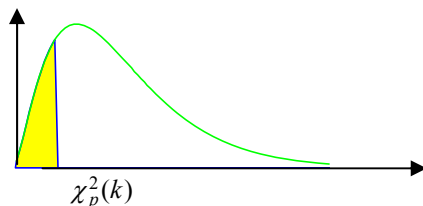
$$p_{\xi}(x) = \begin{cases} \frac{1}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}, & x \geq 0 \\ 0, & x < 0 \end{cases}.$$

На фиг.15.1 са дадени графики при различни стойности на параметъра k . Има таблици за критичните точки или квантилите на тази величина в зависимост от броя k на степените на свобода.

Квантилът от ред p на величина с k степени на свобода ще означаваме с $\chi_p^2(k)$, а стойността му се определяме от таблицата на стр. 170.

Ще отбележим, че има и други означения, а също че в някои литературни източници степените на свобода се означават с df (degree of freedom).

Таблицата за квантилите при дадено k дава стойността $x = \chi_p^2(k)$, за която лицето на заштрихованата област е равна на p (фиг.14.2)



Фиг.14.2.

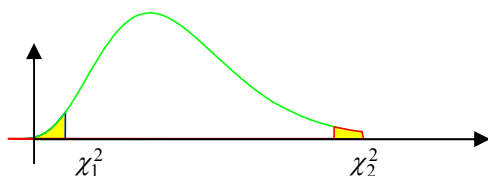
Пример 14.1. Дадена е величината $\xi \sim \chi^2(12)$. Да се намерят тези нейни възможни стойности χ_1^2 и χ_2^2 , за които $P(\xi < \chi_1^2) = P(\xi > \chi_2^2) = 0,025$. Да се намери $P(\chi_1^2 < \xi < \chi_2^2)$

Решение. Очевидно, χ_1^2 е квантилът от ред 0,025 на величината ξ , който намираме от таблицата: $\chi_1^2 = \chi_{0,025}^2(12) = 4,40$.

χ_2^2 е критичната точка от ред 0,025. Имаме

$$P(\xi > \chi_2^2) = 1 - P(\xi \leq \chi_2^2) = 1 - 0,025 = 0,975,$$

т.е. χ_2^2 е квантилът от ред 0,975 $\chi_2^2 = \chi_{0,975}^2(12) = 23,34$ (фиг.15.3)



Фиг.15.3.

Накрая изчисляваме

$$P(\chi_1^2 < \xi < \chi_2^2) = 1 - P(\xi \leq \chi_1^2) - P(\xi \geq \chi_2^2) = 1 - 0,025 - 0,025 = 0,95 \spadesuit$$

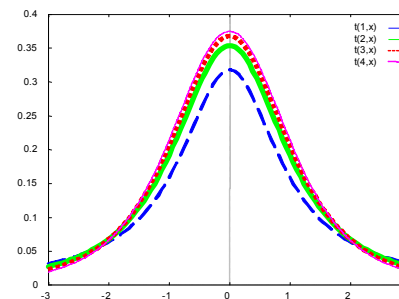
Разпределение на Стьюдент.

Ако $\xi \in N(0,1)$ и η има χ^2 -разпределение с k степни на свобода, то Казваме, че величината $\zeta = \frac{\xi}{\sqrt{\frac{\eta}{k}}}$ има разпределение на Стьюдент (t -разпределение) с k степни на свобода (пишем $\zeta \sim t(k)$).

Плътността на величината $\zeta \sim t(k)$ се получава по формулата

$$p_t(x) = C x^{\frac{k}{2}-1} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}, x \in (-\infty, \infty), \quad (14.1)$$

$$\text{където } C = \frac{1}{\sqrt{k\pi}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \text{ е константа.}$$



Функцията (14.1) е четна, затова графиката ѝ е симетрична относно оста Oy (фиг.14.4).

Частен случай при $k=1$ е разпредението на Коши

$$p(x) = \frac{1}{\pi(1+x^2)} \text{ (пример 9.2).}$$

Квантилът от ред p на F -разпределението ще означаваме с $F_p(k)$, където k са степените

на свобода.

От таблицата на стр. 169 могат да се определят квантилите от редове 0,9, 0,95, и т.н., които намират най-широко приложение.

Пример 14.2. Да се намери симетричен относно началото интервал, в който величината $\xi \sim t(12)$ попада с вероятност 0,9.

Решение. Търсим x , за което $P(-x < \xi < x) = P(|\xi| < x) = 0,9$. Тъй като графиката на функцията е симетрична, то

$$P(\xi < -x) = P(\xi > x) = \frac{1}{2}(1 - 0,9) = 0,05,$$

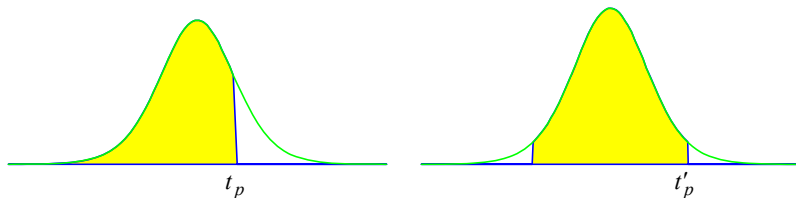
т.е. x е критична точка от ред 0,05. Като имаме предвид връзката между квантил и критична точка, получаваме:

$$x = t_{0,05}^{(kp)}(12) = t_{1-0,05}(12) = t_{0,95}(12) = 1,78.$$

Забележка 14.1. Освен таблица за квантилите на t -разпределението, има и таблици, даващи при дадени p и k вероятността $P(-x < \xi < x) = P(|\xi| < x) = p$.

И двете таблици са съставени от стойностите на x , за които лицето на област под графиката на плътността на разпределение е равно на p , но за първата тази област е дадена на фиг. 14.5а (едностранна област), а за втората – на фиг. 14.5б. (двустранна област). Ако означим а t_p и t'_p числата, получени от двете таблици при една и съща стойност на p , то имаме връзката

$$t'_p = \frac{t_{1+p}}{2}, \quad t_p = \frac{t'_{2p-1}}{2}$$



Фиг.14.5а.

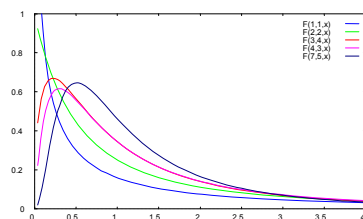
Фиг.14.5б.

Разпределение на Фишер-Снедекор.

Ако $\xi \sim \chi^2(k_1)$ и $\eta \sim \chi^2(k_2)$, то за величината $\zeta = \frac{\xi/k_1}{\eta/k_2}$ казваме, че

има F-разпределение със степени на свобода k_1 и k_2 (разпределение на Фишер-Снедекор), което записваме като $\zeta \sim F(k_1, k_2)$.

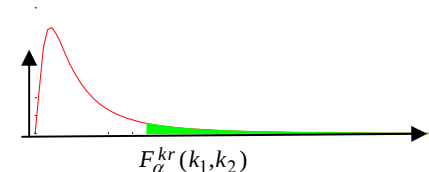
Плътността на тази величина е (фиг 14.6)



Фиг.14.6.

$$p_{\xi}(x) = \begin{cases} C(k_2 + k_1 x)^{-\frac{k_1 + k_2}{2}} x^{\frac{k_1}{2} - 1}, & x \geq 0, \\ 0, & x < 0 \end{cases} \quad C = \frac{\Gamma\left(\frac{k_1 + k_2}{2}\right)}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)} (k_1)^{\frac{k_1}{2}} (k_2)^{\frac{k_2}{2}}.$$

За стойностите $\alpha=0,1, 0.05$ и т.н. при дадени k_1 и k_2 са съставени са таблици за критичните точки $F_{\alpha}^{kr}(k_1, k_2)$ на разпределението (виж стр. 171 и 172), т.е. възможната стойност $F_{\alpha}^{kr}(k_1, k_2)$ на величината $\zeta \sim F(k_1, k_2)$, за която $P(\zeta > F_{\alpha}^{kr}(k_1, k_2)) = \alpha$ (фиг.14.7).



Фиг. 14.7.

Напомниме отново връзката между квантил $F_p(k_1, k_2)$ от ред p и критична точки от ред $1-p$:

$$F_p(k_1, k_2) = F_{1-p}^{kr}(k_1, k_2)$$

Освен това от определението на разпределението на Фишер следва, че ако $\zeta \sim F(k_1, k_2)$ и $\omega \sim F(k_2, k_1)$, то $\zeta = \frac{1}{\omega}$. Тогава, за произволно $x_1 > 0$ събитието $\{\zeta > x_1\}$ е еквивалентно на събитието $\left\{\frac{1}{\omega} > x_1\right\}$, което може да се представи във вида $\left\{\omega < \frac{1}{x_1}\right\}$.

Нека x_1 е критична точка от ред α на $\zeta \sim F(k_1, k_2)$, т.е. $P(\zeta < x_1) = \alpha$. Тогава $P\left(\omega < \frac{1}{x_1}\right) = \alpha$ и, следователно, $\frac{1}{x_1}$ е квантил от ред α на разпределението $\omega \sim F(k_2, k_1)$.

По този начин получаваме следната връзка между квантилите и критичните точки на разпределенията $F(k_1, k_2)$ и $F(k_2, k_1)$:

$$F_{\alpha}(k_1, k_2) = \frac{1}{F_{\alpha}^{kr}(k_2, k_1)}, \quad F_{\alpha}^{kr}(k_1, k_2) = \frac{1}{F_{\alpha}(k_2, k_1)}.$$

НАЙ-ВАЖНИ ДИСКРЕТНИ И НЕПРЕКЪСНАТИ РАЗПРЕДЕЛЕНИЯ

1. Биномно разпределение с параметри n и p : $\xi \sim B(n, p)$, ако възможните и стойности са $k=0,1,\dots,n$ и:

$$P(\xi=k) = p_{\xi}(k) = C_n^k p^k q^{n-k}, \quad E\xi = np, \quad D\xi = npq \quad (q=1-p)$$

Случаи на биномно разпределение: ξ - брой на поява на събитие A в n опита, ако във всеки опит $P(A) = p$.

2. Разпределение на Пуасон с параметър $\lambda > 0$: ($\xi \sim Po(\lambda)$), когато приема изброимо много стойности $0,1,2,\dots,n,\dots$ и

$$P(\xi=k) = p_{\xi}(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k=0,1,\dots,n,\dots, \quad E\xi = \lambda, \quad D\xi = \lambda.$$

Случаи на разпределение на Пуасон:

1) $\xi \sim B(n, p)$ като параметърът е n много голям, а параметърът p – много малък (редки явления), тогава $\xi \sim Po(\lambda)$ като $\lambda = np$.

2) ξ – брой на настъпващите за даден интервал от време елементарни събития при прост поток от събития интензивност λ .

3. Нормално разпределение с параметри a и σ : $\xi \sim N(a, \sigma)$, ако възможните и стойности са $x \in (-\infty, \infty)$ и

$$p_{\xi}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}},$$

$$E\xi = a, \quad D\xi = \sigma^2, \quad \sigma\xi = \sqrt{D\xi} = \sigma \quad a_{\xi} = 0, \quad e_{\xi} = 0$$

$$P(\alpha < \xi < \beta) = F\left(\frac{\beta-a}{\sigma}\right) - F\left(\frac{\alpha-a}{\sigma}\right), \quad P(|\xi - a| < \delta) = 2F\left(\frac{\delta}{\sigma}\right) - 1,$$

където $F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$ е функцията на разпределение на

$$Z \sim N(0, 1).$$

4. Други разпределения с приложения в статистиката:

– χ^2 - разпределение с k степени на свобода: $\xi \sim \chi^2(k)$, ако $\xi = \xi_1^2 + \dots + \xi_k^2$.

– Разпределение на Стюдънт с k степни на свобода: $\zeta \sim t(k)$, ако

$$\zeta = \frac{\xi}{\sqrt{\eta/k}}, \quad \text{където } \xi \in N(0, 1) \text{ и } \eta \in \chi^2(k).$$

– Разпределение на Фишер-Снедекор със степени на свобода k_1 и k_2 :

$$\zeta \sim F(k_1, k_2), \quad \text{ако } \zeta = \frac{\xi}{k_1} : \frac{\eta}{k_2}, \quad \text{където } \xi \sim \chi^2(k_1) \text{ и } \eta \sim \chi^2(k_2).$$

Общи задачи.

1. Какво разпределение има величината и какви са параметрите на разпределението й, ако величината е: а) брой на печатните грешки на страница от книга, ако е известно, че средно на 20 страници има по 1 грешка; б) брой на победите на един спортист в 5 състезания, като вероятността за победа във всяко от тях е 0,8; в) количество лекарство в ампула от 10 мг, ако за произволна ампула количеството лекарство (в мг) е число в интервала (9,91; 10,09). г) продължителност (в години) на експлоатация на уред, ако е установено, че средната продължителност на уреди от този тип е 2 години. д) теглото на стока, измерено без систематическа грешка, ако точността на теглилката е ± 10 г.

2. Дадена е случайната величина $X \sim N(55, 10)$. Да се изчислят а) $P(X > 70)$; б) $P(55 < X < 65)$; в) $P(42 < X < 52)$.

3. Средният брой на грешните свързвания в автоматична телефонна централа е $\lambda = 0,5$. а) Да се напише формулата за плътността на величината ξ – брой на грешните свързвания. Да се намери вероятността да има: а) точно три, б) повече от три грешни свързвания.

4. Съобщение, се състои от 3 символа. Всеки символ, независимо от останалите се изкривява с вероятност 0,1. а) Каква е вероятността съобщението да бъде прието без изкривяване. За по-голяма надеждост, съобщението се предава 2 пъти. б) Да се състави законът на величината ξ – брой на вярно предадените съобщения и намерят математическото очакване и дисперсията й. в) Каква е вероятността поне едно от съобщенията да е прието без изкривяване?

5. Теглото на жителите в дадена област е нормално разпределена случайна величина със средна стойност 80 кг и средно отклонение 8кг. а) Да се скицира графиката на плътността на теглото X на жител от областта. б) Какъв е процентът на жителите, които са по-тежки от 90 кг? в) Каква е вероятността случайно избран жител да е по-лек от 60 кг?

6. Средният брой съобщения за 1 час по електронната поща е 0,6. а) Каква е вероятността за 1 час да има повече от 1 съобщения? б) Пощата е проверявана в 10, 11, 12 и 13 часа. Каква е вероятността да не е получено нито едно съобщение?

7. Времето за извършване на дадена работа е нормално разпределена случайна величина с математическо очакване 60 минути и средно квадратично отклонение 30 минути. Каква е вероятността че произволно избран работник: а) ще завърши работата за по-малко от 30 минути; б) за какво време ще извършат работата 99% от работниците?

8. Случайната величина ξ има разпределение на Поасон с математическо очакване $E\xi = 1$. Да се намери: а) $P(\xi < 2)$; б) $P(\xi > 2)$.

9. Известно е, че 20% от студентите работят. а) Да се състави таблицата на разпределение на величината брой на работещите сред 3 произволно избрани студента. б) Да се намерят средната и дисперсията на извадката.

10. Дадена е случайната величина $\xi \sim B(8, 1/3)$. Да се намери: а) $P(\xi \leq 2)$, б) $P(\xi \geq 2)$.

11. В аквариум има 2 големи и 3 малки рибки. Ако се хване малка рибка, тя се пуска обратно в аквариума, ако се хване голяма – се продава. Колко средно опита трябва да се извършат, че в аквариума да останат само малки рибки?

12. Отборът А печели кой да е мач с вероятност $2/3$. Каква е вероятността от 4 мача отборът да спечели не по-малко от половината?

13. Ако $\xi \sim B(10, 0,4)$ и $\eta \sim N(0, 1)$, какви са математическото очакване и дисперсията на величината $\xi + \eta$.

14. За да се спечели награда, трябва да се съберат 5 различни картинки, поставени по случаен начин във всеки пакет детски бонбони. Средно колко пакета трябва да се закупят, за да се съберат десетте различни картинки?

15. Да се изчисли $P(\xi < 0,95)$, ако: а) $\xi \sim \chi^2(12)$; б) $\xi \sim t(13)$; в) $\xi \sim F(8, 10)$.

16. В кутия има 4 червени и 5 черни топки. Изважда се една топка и ако е червена, се връща обратно, ако е черна се отстранява. Да се намери колко е средният брой тегления докато се отстранят всички черни топки.

§15. Гранични теореми.

В §13 получихме разпределението на сумата на няколко величини с еднакво показателно разпределение. При изследване на графиките на плътностите им в зависимост от броя на събираемите се забелязва изместване на графиките надясно и все по-голямата прилика с графиката на нормално разпределена величина. Сега ще разгледаме една важна теорема, наречена централна гранична теорема, която обяснява това явление и която доказва фундаменталната роля на величините с нормално разпределение.

Да разгледаме редицата $\xi_1, \xi_2, \dots, \xi_n, \dots$ от независими случайни величини с **еднакво разпределение**, които имат математическо очакване μ и средно квадратично отклонение σ . Образоваме величината $\xi = \xi_1 + \xi_2 + \dots + \xi_n + \dots$. От свойствата на математическото очакване и дисперсията (виж §8) получаваме

$$E\xi = E(\xi_1 + \xi_2 + \dots + \xi_n) = E\xi_1 + E\xi_2 + \dots + E\xi_n = n\mu, \quad (15.1)$$

$$D\xi = D(\xi_1 + \xi_2 + \dots + \xi_n) = D\xi_1 + D\xi_2 + \dots + D\xi_n = n\sigma^2, \quad (15.2)$$

От тук следва, че за величината $Z = \frac{\xi_1 + \dots + \xi_n - n\mu}{\sigma\sqrt{n}}$ имаме

$$EZ = \frac{1}{\sigma\sqrt{n}}(n\mu - n\mu) = 0, \text{ т.е. } Z \text{ е центрирана и дисперсията ѝ е}$$

$$DZ = \left(\frac{1}{\sigma\sqrt{n}}\right)^2 D(\xi_1 + \dots + \xi_n - n\mu) = \frac{D\xi_1 + \dots + D\xi_n - 0}{\sigma^2 n} = \frac{\sigma^2 n}{\sigma^2 n} = 1.$$

Теоремата, която ще изкажем гласи, че с нарастването на броя на събираемите законът на величината Z , за която показахме, че $EZ=0$, $DZ=1$, се доближава до закона на нормално разпределена величина с същите математическо очакване и средно квадратично отклонение, т.е. $Z \sim N(0,1)$.

Теорема 15.1 (Централна гранична теорема). Ако $\xi_1, \xi_2, \dots, \xi_n, \dots$ са независими случайни величини с еднакво разпределение, които имат математическо очакване μ и средно квадратично отклонение σ и

$$Z = \frac{\xi_1 + \dots + \xi_n - n\mu}{\sigma\sqrt{n}}, \text{ то за произволни } \alpha < \beta$$

$$\lim_{n \rightarrow \infty} P(\alpha < Z < \beta) = \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-\frac{x^2}{2}} dx.$$

Забележка 15.1. Интегралът в последната формула е вероятността стандартна нормална величина да приеме стойност в интервала (α, β) , следователно, Z е величина с разпределение, близко до $N(0,1)$.

Забележка 15.2. Случайната величина Z може да се представи във вида (делим числителя и знаменателя на n)

$$Z = \frac{\xi_1 + \dots + \xi_n - n\mu}{\sigma\sqrt{n}} = \frac{\frac{1}{n}(\xi_1 + \dots + \xi_n) - EX}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{\xi} - \mu}{\frac{\sigma}{\sqrt{n}}},$$

където $\bar{\xi} = \frac{\xi_1 + \dots + \xi_n}{n}$ е средно аритметичното на величините $\xi_1, \xi_2, \dots, \xi_n$.

Като имаме предвид, че от формули (15.1) и (15.2) следва:

$$\begin{aligned} E\bar{\xi} &= \frac{1}{n}E(\xi_1 + \dots + \xi_n) = \frac{n\mu}{n} = \mu \\ D\bar{\xi} &= \frac{1}{n^2}D(\xi_1 + \dots + \xi_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}, \quad \sigma_{\bar{\xi}} = \frac{\sigma}{\sqrt{n}} \end{aligned} \quad (15.3)$$

може да обобщим:

Ако $\xi_1, \xi_2, \dots, \xi_n$ са независими случайни величини с еднакво разпределение, $E\xi_i = \mu$, $D\xi_i = \sigma^2$, $i=1, \dots, n$, то:

- Сумата им $\xi = \xi_1 + \dots + \xi_n$ е случайна величина, за която $E\xi = n\mu$ и $\sigma_{\xi} = \sqrt{D\xi} = \sqrt{n\sigma^2} = \sigma\sqrt{n}$. При големи стойности на n разпределението на ξ е близко до разпределението $N(n\mu, \sigma\sqrt{n})$;

- Средноаритметичното им $\bar{\xi} = \frac{\xi_1 + \dots + \xi_n}{n}$ е случайна величина, за която $E\bar{\xi} = \mu$ и $\sigma_{\bar{\xi}} = \frac{\sigma}{\sqrt{n}}$. При големи стойности на n разпределението на $\bar{\xi}$ е близко до разпределението $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Пример 15.1. Автомат произвежда детайли, дължината на които се контролира и която е нормално разпределена случайна величина ξ с математическо очакване 3,5 и средно квадратично отклонение 0,1. Избрани са случайно 100 детайла (случайна извадка с обем 100) и е намерено средното им аритметично $\bar{\xi}$. Да се определи какъв е законът на разпределение на случайната величина $\bar{\xi}$ и намерят математическото очакване и средно квадратичното ѝ отклонение.

Решение. Въвеждаме величините ξ_i - дължината на i -тия детайл, за които $\xi_i \sim N(3,5; 0,1)$. Тъй като величините са независими и имат еднакво разпределение, то по формула (15.3) имаме

$$E\bar{\xi} = 3,5, \quad D\bar{\xi} = \frac{0,1^2}{100}, \quad \sigma_{\bar{\xi}} = \frac{0,1}{10} = 0,01.$$

Освен това, както гласи теорема **13.1**, величината $\bar{\xi}$ е нормално разпределена. Следователно, $\bar{\xi} \sim N(3,5; 0,01)$, т.е. *средно аритметичното на величини с еднакво нормално разпределение е нормално разпределено със същото математическо очакване и с $\sqrt{n}$ пъти по-малко средно квадратично отклонение.* ♦

Условията за прилагане на теорема **15.1** са твърде общи – иска се само 1) *величините да са независими*, 2) *да имат едно и също разпределение* и 3) *да имат математическо очакване и дисперсия*. А много често една случайна величина приема една или друга стойност именно под действието на множество други случайни и независими събития и затова се представя като сума от множество случайни величини с еднакви разпределения и приемащи малки стойности. Например:

- Нalleyгането на газа върху стените на съда се дължи на бомбардирането му от молекулите на газа (на брой около 10^{22} за 1 литър), което е зависи от скоростта на молекулите.
- Смутенията в много технически системи се дължат на голям брой малки независими случайни фактори и опитно е установено, че характерът на смутенията се описва с нормален закон за разпределение.

От гледна точка на практиката теоремата *премахва необходимостта от прецизен вероятностен модел* и дава възможност пресмятането на вероятност да се сведе до прилагане на формули и таблици за нормалното разпределение.

Пример 15.2. През последните години са регистрирани средно по 49 рекламации годишно. Каква е вероятността че тази година броя рекламациите да бъде: а) 30: б) между 47 и 50.

Решение. Величината ξ – брой на рекламациите за една година е поасоново разпределена с параметър $\lambda=49$ (среден брой рекламации за година). Тъй като $\lambda=49$ е голямо число (e^{-49} е много малко), се използва като приближение величина η с нормално разпределение, за която $E\eta=E\xi=\lambda=49$ и $\sigma\eta=\sigma\xi=\sqrt{\lambda}=\sqrt{49}=7$. Тъй като величината ξ е дискретна, а величината $\eta \sim N(49,7)$ е непрекъсната, то един от начините за пресмятане на вероятности на събития, касаещи ξ , е следният:

$$\text{а) } P(\xi=30)=P(29,5 \leq \eta \leq 31,5)=F\left(\frac{31,5-49}{7}\right)-F\left(\frac{29,5-49}{7}\right)=1-F(2,5)-1+F(2,78)=0,9973-0,9940=0,0033$$

$$\text{б) } P(47 < \xi < 50)=P(47,5 \leq \eta \leq 49,5)=F\left(\frac{49,5-49}{7}\right)-F\left(\frac{47,5-49}{7}\right)=F(0,7)-1+F(0,21)=0,7580-1+0,5832=0,3412. \quad \blacklozenge$$

Ще приложим теорема **15.1** в следния частен случай. Нека $P(A)=p$ във всеки опит (т.е. извършваме опити по схемата на Бернули) и нека ξ_i е индикаторът на случайното събитие A за i -тия опит. Както знаем, за тези величини имаме

$$\xi_i \begin{cases} 0 & 1 \\ p & q \end{cases}, \quad E\xi_i=p, \quad D\xi_i=q(0-p)^2+p(1-p)^2=qp(q+p)=qp,$$

а сумата им $\xi=\xi_1+\dots+\xi_n$ приема стойност, равна на броя k на настъпване на A в n опита. При това $E\xi=np$, $\sigma\xi=\sqrt{npq}$ (виж §11 за биомно разпределение). Като вземем под внимание следствие **15.1**, получаваме че $\xi \sim N(np, \sqrt{npq})$, откъдето следва

Теорема 15.2 (Централна (интегрална) гранична теорема на Моавър-Лаплас) Ако $\xi \sim B(n,p)$ е биомно разпределена величина, то при големи стойности на n вероятността ξ да приеме стойност в интервала $[a,b]$ се пресмята по формулата

$$P(a \leq \xi \leq b) \approx F\left(\frac{b-np}{\sqrt{npq}}\right) - F\left(\frac{a-np}{\sqrt{npq}}\right). \quad (15.4)$$

Следователно, при големи стойности на n разпределението на биомна случайна величина се доближава до разпределението на нормално разпределена величина със същото математическо очакване и дисперсия. От тук следва и формулата

$$P_n(k) = C_n^k p^k q^{n-k} \approx \frac{1}{\sqrt{2\pi npq}} e^{-\frac{(k-np)^2}{2npq}},$$

която се доказва в така наречената **локална гранична теорема на Моавър-Лаплас**.

Пример 15.3. Нека ξ е биомно разпределена величина с параметри $p=0,5$ и $n=36$. Ще сравним стойността за $P(\xi \leq 21)$, получена:

$$P(\xi \leq 21) = \sum_{k=0}^{21} C_{36}^k p^k q^{36-k} = \sum_{k=0}^{21} \binom{36}{k} (0,5)^{36} = 0,8785 \text{ - точно пересмятане,}$$

$$\text{По формула (15.3): } P(\xi \leq 21) = F\left(\frac{21-36 \cdot 0,5}{\sqrt{36 \cdot 0,5 \cdot 0,5}}\right) = 0,8413.$$

Пример 15.4. Селище A има 2500 жители. Всеки от тях, примерно, 6 пъти месечно пътува до селище B по случайни, независещи от останалите жители причини. Колко места трябва да има влакът от A до B ,

че той да е препълнен не повече от 1 на 1000 дни, ако се движи веднъж дневно.

Решение. Нека $C=\{\text{жител пътува за селище } B\}$. $P(C)=\frac{6}{30}=0,2$.

Величината ξ - брой на пътуващите за B на ден е биомно разпределена с параметри $p=0,2$ и $n=2500$. Трябва да се определи броят N на местата във вагоните така, че $P(\xi > N) \leq 0,001$. Прилагаме формулата на

Моавър-Лаплас: $P(N < \xi < \infty) = 1 - F\left(\frac{N-2500,0,2}{\sqrt{2500,0,2,0,8}}\right) = 1 - F\left(\frac{N-500}{20}\right) \leq 0,001$

$\Rightarrow F\left(\frac{N-500}{20}\right) \geq 0,9999$. Тъй като $F(4)=0,9999$, то $\frac{N-500}{20} \geq 4 \Rightarrow N \geq 580$.

Следователно, необходимо е влакът да има 580 места. ♦

Забележка 15.3. За биомното разпределение може да използваме две приближения:

при голямо n и малко p - разпределение на Поасон $Po(np)$.

при голямо n и $np > 5$ - нормално разпределение $N(np, \sqrt{npq})$.

Упражнения.

1. Да се сравнят намерените вероятности в задача 7, §9, с приближените им стойности, пресметнати по формулите на Моавър-Лаплас и на Поасон.

2. Да се намери вероятността от 100 хвърляния да се падне 50 пъти лицето на монетата.

3. Нека $\xi_1, \dots, \xi_{25}$ са 25 независими случайни величини с експоненциално разпределение с параметър $\lambda=2$. Разглеждаме величината $\bar{\xi} = \frac{1}{25}(\xi_1 + \dots + \xi_{25})$.

Да се изчисли $P(\bar{\xi} < 0,25)$.

4. Частица еднакво вероятно попада в произволна точка от интервала $[-10,10]$. Каква е вероятността средното разстояние от началото до мястото на падане на 100 частици да е по-голямо от 1?

5. Малка частица скача с еднаква вероятност наляво и надясно по права линия. Всеки скок е с дължина 1 мм и се извършва за 1 милисек. а) След 10 секунди каква е вероятността частицата да е на разстояние от началното си положение по-голямо от 5 см. б) След колко време с вероятност 0,5 частицата се намира по-далеч от 5 см?

6. Случайната величина ξ приема стойностите 1, 2 и 3 с вероятности съответно 0,25, 0,5, 0,25. Каква е вероятността средно аритметичното от 72 независими измервания на ξ да е по-малко от 2,1?

§16. Неравенство на Чебишев. Закон за големите числа.

В началото бе отбелязан фактът, че при дълги серии от опити относителната честота на поява на случайно събитие се колебае около определено число, затова то се приема като мярка за това колко често

настъпва събитието (виж §3 - статистическа вероятност). Сега ще докажем това свойство на относителната честота като по този начин се убеждаваме, че избраните математически модели на случаен експеримент и на вероятност на събитие се съгласуват с действителността. Ще обосновем също, защо при достатъчно много опити може да твърдим, че средно аритметичното на получените стойности на изследваната случайна величина не се различава съществено от математическото ѝ очакване.

Групата теореми, които ще разгледаме и които установяват и обясняват устойчивостта на средните стойности на голям клас случайни величини, се обединяват под името закон за големите числа.

Теорема 16.1 (неравенство на Чебишев). Ако случайната величина ξ има дисперсия и $\varepsilon > 0$ е произволно число, то в сила е неравенството

$$P(|\xi - E\xi| < \varepsilon) \geq 1 - \frac{D\xi}{\varepsilon^2}. \quad (16.1)$$

Доказателство. Очевидно, неравенството на Чебишев може да се запише във вида

$$P(|\xi - E\xi| \geq \varepsilon) \leq \frac{D\xi}{\varepsilon^2}. \quad (16.2)$$

Ще докажем последното неравенство за дискретна случайна величина като оценим дисперсията ѝ. Съгласно формула (8.2) представяме $D\xi$ във вида ($p_i = p_\xi(x_i)$)

$$D\xi = \sum_i (x_i - E\xi)^2 p_i = \sum_{x_i: |x_i - E\xi| < \varepsilon} (x_i - E\xi)^2 p_i + \sum_{x_i: |x_i - E\xi| \geq \varepsilon} (x_i - E\xi)^2 p_i.$$

Първата сума е неотрицателна, а за всяко събираемо на втората имаме

$$|x_i - E\xi| \geq \varepsilon, \text{ откъдето } \sum_{x_i: |x_i - E\xi| \geq \varepsilon} (x_i - E\xi)^2 p_i \geq \varepsilon^2 \left(\sum_{x_i: |x_i - E\xi| \geq \varepsilon} p_i \right).$$

$$\text{Но } \sum_{x_i: |x_i - E\xi| \geq \varepsilon} p_i = P(|\xi - E\xi| \geq \varepsilon). \text{ Следователно, } D\xi \geq \varepsilon^2 P(|\xi - E\xi| \geq \varepsilon).$$

$$\text{От тук } P(|\xi - E\xi| \geq \varepsilon) \leq \frac{D\xi}{\varepsilon^2} \Rightarrow P(|\xi - E\xi| < \varepsilon) \geq 1 - \frac{D\xi}{\varepsilon^2} \quad \diamond$$

Пример 16.1. Вероятността за настъпване на събитието A във всеки от 1500 опита е 0,2. Нека ξ е броят на поява събитието. Използвайки неравенството на Чебишев, да се оцени вероятността за това, че отклонението на величината ξ от математическото ѝ очакване да бъде по-голямо от 40.

Решение. Величината ξ има биомно разпределение с математическо очакване $E\xi = 1500 \cdot 0,2 = 300$ и дисперсия $D\xi = 240$.

Прилагаме (16.1) за $\varepsilon=40$: $P(|\xi-300|<40)\geq 1-\frac{240}{40^2}=0,15$. ♦

Пример 16.2. При подхвърляне на монета 1000 пъти наблюдаваме събитието $A=\{\text{поява на герб}\}$. Да се оцени вероятността отклонението на относителната честота $w(A)$ от негова вероятност $P(A)$ да бъде по-малко от 0,1.

Решение. Ако ξ_i е индикатора на A в i -тия опит, то $\xi = \frac{\xi_1 + \dots + \xi_n}{n} = \frac{m}{n}$ е относителната честота на A в n опита. За да приложим неравенството, пресмятаме

$$E\xi = \frac{1}{n}(E\xi_1 + \dots + E\xi_n) = \frac{1}{n} \cdot n \cdot \frac{1}{2} = \frac{1}{2}, \quad D\xi = \frac{1}{n^2}(D\xi_1 + \dots + D\xi_n) = \frac{1}{n^2} \cdot n \cdot \frac{1}{4} = \frac{1}{4n}.$$

При $n=1000$ получаваме $P\left(\left|\frac{m}{1000} - \frac{1}{2}\right| < 0,1\right) \geq 1 - \frac{1}{4000 \cdot 0,01} = \frac{39}{40}$. ♦

Неравенство (16.1) оценява вероятността за отклоняване на случайната величина от математическото ѝ очакване. Но не бива да се преувеличава приложението на (16.1) и (16.2), тъй като тези оценки са доста груби. Например, ако $\varepsilon^2 < D\xi$, то неравенство (16.1) дава $P(|\xi - E\xi| < \varepsilon) > 0$, което е предварително известно. Главното приложение на (16.1) е, че с него се доказва теорема, наричана закон за големите числа. Тази теорема “действа” в много физични явления. Например, знаем, че да се предскаже поведението на една от частиците на газ е невъзможно, но сумарното поведение на всички частици е предсказуемо, Смисълът на теоремата е именно в това – да се предскаже стойността, която ще приеме една от много случайни величини $\xi_1, \xi_2, \dots, \xi_n$, е невъзможно, но средноаритметичното от стойностите им е относително постоянно и това е изпълнено за широк клас случайни величини. Теоремата гласи:

Теорема 16.2. Ако случайните величини $\xi_1, \xi_2, \dots, \xi_n, \dots$ са независими помежду си и дисперсиите им са ограничени от едно и също число C , т.е. $D\xi_1 \leq C, \dots, D\xi_n \leq C$, то за всяко $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{\xi_1 + \xi_2 + \dots + \xi_n}{n} - \frac{E\xi_1 + E\xi_2 + \dots + E\xi_n}{n}\right| < \varepsilon\right) = 1.$$

Доказателство. Разглеждаме величината $\bar{\xi} = \frac{\xi_1 + \dots + \xi_n}{n}$ и за която пресмятаме математическото очакване и дисперсията

$$E\bar{\xi} = \frac{E\xi_1 + E\xi_2 + \dots + E\xi_n}{n},$$

$$D\bar{\xi} = \frac{1}{n^2}(D\xi_1 + \dots + D\xi_n) \leq \frac{1}{n^2} \cdot n \cdot C = \frac{C}{n},$$

Прилагаме неравенството на Чебишев (16.1), т.е. $P(|\bar{\xi} - E\bar{\xi}| < \varepsilon) \geq 1 - \frac{D\bar{\xi}}{\varepsilon^2}$.

Следователно,

$$P\left(\left|\frac{\xi_1 + \xi_2 + \dots + \xi_n}{n} - \frac{E\xi_1 + E\xi_2 + \dots + E\xi_n}{n}\right| < \varepsilon\right) \geq 1 - \frac{D\bar{\xi}}{\varepsilon^2} \geq 1 - \frac{C}{n\varepsilon^2}.$$

Равенството се получава при $n \rightarrow \infty$. ♦

Ще приложим теоремата в случаите:

Теорема 16.3. (Чебишев) Ако величините $\xi_1, \xi_2, \dots, \xi_n, \dots$ са еднакво разпределени, независими са помежду си и имат крайна дисперсия, то за всяко $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{\xi_1 + \xi_2 + \dots + \xi_n}{n} - E\xi\right| < \varepsilon\right) = 1,$$

където $E\xi$ е математическото очакване на величините.

Доказателството следва от това, че в случая $E\xi_1 = \dots = E\xi_n = E\xi$.

Тази теорема служи за обосноваване на основния принцип на математическата статистика, според който по случайна извадка на изучаваната съвкупност могат да се правят изводи за цялата съвкупност от изследвани обекти.

Да приложим теорема 16.3 за схемата на Бернули. Нека в серия от n опита наблюдаваме броя на поява на събитието A с една и съща вероятност за настъпване $P(A) = p$. Нека ξ_k е индикаторът на събитието A за k -тия опит. Тогава $\frac{\xi_1 + \xi_2 + \dots + \xi_n}{n} = \frac{m}{n}$, където m е броят на поява на събитието A в серията. От друга страна $\xi_1, \xi_2, \dots, \xi_n$ са еднакво разпределени величини като $E\xi_k = p$ (виж §11). Като приложим теорема 16.3 получаваме твърдението

Теорема 16.4. (Бернули) Нека m е броят на поява на A в n опита и $P(A) = p$ във всеки опит. Тогава, ако $\varepsilon > 0$, то

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{m}{n} - p\right| < \varepsilon\right) = 1$$

С тази теорема се обосновава устойчивото поведение на относителната честота $w(A) = \frac{m}{n}$ на случайно събитие.

ДВУМЕРНИ СЛУЧАЙНИ ВЕЛИЧИНИ

§17. Двумерна случайна величина - закон за разпределение. Закони за разпределение на съставните величини.

В теорията на вероятностите често се разглеждат съвместно няколко случайни величини. Например, при изследване на няколко признака на някакво физическо явление, въздействието на дадено лекарство върху различни функции на човешкия организъм и т.н. Едни от основните въпроси са:

- какви са законите на разпределение на тези величини;
- дали величините са зависими едни от други;
- какъв е характерът на зависимостта между тях.

Нека $\xi_1, \xi_2, \dots, \xi_n$ са случайни величини. Величината $(\xi_1, \xi_2, \dots, \xi_n)$ се нарича *n*-мерна случайна величина, а величините $\xi_1, \xi_2, \dots, \xi_n$ - нейни съставни.

Геометрически *n*-мерната случайна величина се интерпретира като случайна точка в *n*-мерното пространство.

Ще разгледаме частния случай на двумерна случайна величина (ξ, η) . Областта на възможните стойности на величината (ξ, η) е област от координатната равнина Oxy , а за да бъде величината (ξ, η) определена, трябва можем да намираме вероятността $P((\xi, \eta) \in D)$, за произволна двумерна област D .

Закон за разпределение на вероятностите на (ξ, η) (съвместен закон на разпределение на съставните ξ и η) се нарича всяко правило, с помощта на което може да се намери вероятността на произволно случайно събитие, отнасящо се за тази величина.

Дискретни двумерни величини.

Ако ξ и η са дискретни, възможните стойности на които са x_i и y_j , $i=1, 2, \dots, n$, $j=1, 2, \dots, m$, то величината (ξ, η) се нарича дискретна.

Законът на разпределението на дискретна двумерна величина най-често се задава с плътността на разпределението – функция $p_{\xi, \eta}(x_i, y_j)$, дефинирана за възможните стойности на величината и установяваща връзка между възможните стойности (x_i, y_j) и вероятността p_{ij} за приемането им, а именно:

$$p_{\xi, \eta}(x_i, y_j) = P(\xi = x_i \cap \eta = y_j) = p_{ij}.$$

Събитията $\{\xi = x_i \cap \eta = y_j\}$ образуват пълна група събития, затова

$$\sum_i \sum_j p_{ij} = 1.$$

Обикновено възможните стойности и вероятностите p_{ij} се представят с таблица от вида (таблица на разпределението)

ξ	x_1	$\dots$	x_n	
η	y_1	p_{11}	$\dots$	p_{n1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_m	p_{1m}	$\dots$	p_{nm}	

$$\sum_i \sum_j p_{ij} = 1$$

Сумата $\sum_{j=1}^m p_{ij}$ е безусловната вероятност $p_{\xi}(x_i)$ на събитието

$\{\xi = x_i\}$, тъй като не зависи от това каква стойност е приела величината η . По този начин, събирайки вероятностите p_{ij} по стълбове получаваме безусловния закон на разпределение на съставната случайна величина ξ . Ако съберем вероятностите по редове, получаваме безусловния закон на величината η . Получените таблици на разпределение имат вида

$$\begin{array}{c|ccc} \xi & x_1 & \dots & x_n \\ \hline P & p_{\xi}(x_1) & \dots & p_{\xi}(x_n) \end{array} \quad \begin{array}{c|ccc} \eta & y_1 & \dots & y_m \\ \hline P & p_{\eta}(y_1) & \dots & p_{\eta}(y_m) \end{array}.$$

Пример 17.1. Зарче се подхвърля един път. Разглеждат се събитията $A = \{\text{пада се четно число}\}$ с индикатор случайната величина ξ и $B = \{\text{пада се число, кратно на 3}\}$ с индикатор случайната величина η . Да се състави таблицата на разпределение на величината (ξ, η) . Да се намерят законите на разпределение на величините ξ и η .

Решение. Например, $\xi=0$ и $\eta=0$ означава, че са се паднали 1 или 5 точки, следователно, $P(\xi=0 \cap \eta=0) = \frac{2}{6} = \frac{1}{3}$. По аналогичен начин изчисляваме и другите вероятности. Така получаваме (табл.18.1)

ξ	0	1	$P_{\eta}(y_i)$
η			
0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$
1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$
$P_{\xi}(x_i)$	$\frac{1}{2}$	$\frac{1}{2}$	1

В тази таблица, освен съвместното разпределение на двете величини ξ и η са получени и безусловните им закони – последният стълб (получен чрез събиране по редове) е съставен от вероятностите $P(\eta = y_j) = p_{\eta}(y_j)$, а последният ред (получен чрез събиране по стълбове) – от вероятностите $P(\xi = x_i) = p_{\xi}(x_i)$. ♦

Непрекъснати двумерни величини.

Непрекъснати случайни величини (ξ, η) са тези, за които съставните ξ и η са непрекъснати. Те се задават с помощта на функцията плътност $p_{\xi, \eta}(x, y)$ на съвместното разпределение, дефинирана в цялата равнина Oxy , която има свойствата:

$$1) p_{\xi, \eta}(x, y) \geq 0, (x, y) \in R^2, \quad 2) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{\xi, \eta}(x, y) dx dy = 1;$$

$$3) P((x, y) \in G) = \iint_G p_{\xi, \eta}(x, y) dx dy.$$

Плътностите на съставните ξ и η се намират по формулите:

$$p_{\xi}(x) = \int_{-\infty}^{\infty} p_{\xi, \eta}(x, y) dy \text{ -плътност на } \xi,$$

$$p_{\eta}(y) = \int_{-\infty}^{\infty} p_{\xi, \eta}(x, y) dx \text{ -плътност на } \eta,$$

И тук се дефинира

функция на разпределението: $F_{\xi, \eta}(x, y) = P(\xi < x \cap \eta < y), (x, y) \in R$.

За дискретни случайни величини функцията на разпределение е прекъсната във всяка точка $(x_i, y_j), i=1, \dots, n, j=1, \dots, m$.

За непрекъснати случайни величини предполагаме, че функцията $F_{\xi, \eta}(x, y)$ е непрекъсната. Имаме следните връзки между плътността и функцията на разпределение на двумерна случайна величина:

$$p_{\xi, \eta}(x, y) = \frac{\partial^2 F}{\partial x \partial y}, \quad F_{\xi, \eta}(x, y) = \int_{-\infty}^x \int_{-\infty}^y p_{\xi, \eta}(x, y) dx dy.$$

При зададена функция $F_{\xi, \eta}(x, y)$ функцията на разпределение на компонентата ξ на непрекъснатата величина (ξ, η) , може да намерим чрез граничния преход $F_{\xi}(x) = \lim_{y \rightarrow \infty} F_{\xi, \eta}(x, y)$.

Допълвайки с останалите свойства на функцията на разпределение на двумерна случайна величина, получаваме:

$$\lim_{x \rightarrow -\infty} F_{\xi, \eta}(x, y) = \lim_{y \rightarrow -\infty} F_{\xi, \eta}(x, y) = \lim_{x \rightarrow -\infty, y \rightarrow -\infty} F_{\xi, \eta}(x, y) = 0,$$

$$\lim_{x \rightarrow \infty} F_{\xi, \eta}(x, y) = F_{\eta}(y), \quad \lim_{y \rightarrow \infty} F_{\xi, \eta}(x, y) = F_{\xi}(x), \quad \lim_{x \rightarrow \infty, y \rightarrow \infty} F_{\xi, \eta}(x, y) = 1$$

$$P(a \leq \xi < b, c \leq \eta < d) = F_{\xi, \eta}(b, d) - F_{\xi, \eta}(a, d) - [F_{\xi, \eta}(b, c) - F_{\xi, \eta}(a, c)].$$

Може да обобщим:

Безусловни закони на съставните величини

$$p_{\xi}(x_i) = \sum_j p_{ij} \text{ и } p_{\eta}(y_j) = \sum_i p_{ij}, \text{ ако } (\xi, \eta) \text{ е дискретна.}$$

$$p_{\xi}(x) = \int_{-\infty}^{\infty} p_{\xi, \eta}(x, y) dy \text{ и } p_{\eta}(y) = \int_{-\infty}^{\infty} p_{\xi, \eta}(x, y) dx, \text{ ако } (\xi, \eta) \text{ е непрекъсната}$$

От тези безусловни закони може да получим всички числени характеристики на ξ и η като $E\xi$ и $E\eta$, $D\xi$ и $D\eta$ и т.н. Например,

$$E\xi = \begin{cases} \sum_i p_{\xi}(x_i) x_i, & \xi - \text{дискр.} \\ \int_{-\infty}^{\infty} p_{\xi}(x) x dx, & \xi - \text{непр.} \end{cases}, \quad E\eta = \begin{cases} \sum_i p_{\eta}(y_i) y_i, & \eta - \text{дискр.} \\ \int_{-\infty}^{\infty} p_{\eta}(y) y dy, & \eta - \text{непр.} \end{cases},$$

Ще отбележим също и следното свойство на независимите случайни величини:

$$\text{Величините } \xi \text{ и } \eta \text{ са независими} \Leftrightarrow \begin{aligned} p_{\xi, \eta}(x, y) &= p_{\xi}(x) p_{\eta}(y) \\ F_{\xi, \eta}(x, y) &= F_{\xi}(x) F_{\eta}(y) \end{aligned}$$

Пример 17.2. Да се намерят безусловните закони на величините, ξ и η , ако е дадена плътността на съвместното им разпределение $p_{\xi, \eta}(x, y) = 6x^2y$, за $0 < x < 1, 0 < y < 1$ и $p_{\xi, \eta}(x, y) = 0$ във всички останали случаи.

Решение. Съгласно формулите

$$p_{\xi}(x) = \int_{-\infty}^{\infty} p_{\xi, \eta}(x, y) dy = 6 \int_0^1 x^2 y dy = 3x^2 \text{ за } x \in (0, 1) \text{ и } p_{\xi}(x) = 0 \text{ за } x \notin (0, 1).$$

$$p_{\eta}(y) = \int_{-\infty}^{\infty} p_{\xi, \eta}(x, y) dx = 6 \int_0^1 x^2 y dx = 2y \text{ за } y \in (0, 1) \text{ и } p_{\eta}(y) = 0 \text{ за } y \notin (0, 1).$$

Очевидно, $p_{\xi}(x) p_{\eta}(y) = 3x^2 \cdot 2y = 6x^2y = p_{\xi, \eta}(x, y)$, откъдето получаваме, че величините ξ и η са независими. ♦

Пример 17.3. Нека $p_{\xi,\eta}(x,y)=x+y$ за $(0<x<1, 0<y<1)$ и $p_{\xi,\eta}(x,y)=0$ в останалите случаи. Да се докаже, че величините ξ и η са зависими.

Решение. Безусловните закони на величините са

$$p_{\xi}(x)=\int_0^1(x+y)dy=x+\frac{1}{2}, \quad x \in (0,1), \quad p_{\eta}(y)=\int_0^1(x+y)dx=y+\frac{1}{2}, \quad y \in (0,1).$$

Очевидно, $p_{\xi,\eta}(x,y) \neq p_{\xi}(x) \cdot p_{\eta}(y)$, т.е. величините са зависими. ♦

Пример 17.4. За величината (ξ, η) със закон на разпределение

$\eta \backslash \xi$	-1	0	1
1	0,15	0,3	0,35
2	0,05	0,05	0,1

да се намерят $E\xi$, $E\eta$, $D\xi$ и $D\eta$.

Решение. Намираме безусловните закони (може да допълним таблицата на съвместното разпределение както в пример 17.1.

ξ	-1	0	1
P	0,2	0,35	0,45

η	1	2
P	0,8	0,2

откъдето: $E\xi = -0,2 + 0,45 = 0,25$, $E\eta = 1 \cdot 0,8 + 2 \cdot 0,2 = 1,2$

$$E\xi^2 = (-1)^2 \cdot 0,2 + 1^2 \cdot 0,45 = 0,65, \quad E\eta^2 = 1^2 \cdot 0,8 + 2^2 \cdot 0,2 = 1,6$$

$$D\xi = 0,65 - 0,25^2 = 0,5875 \quad D\eta = 1,6 - 1,2^2 = 0,16. \quad \spadesuit$$

Сума на две случайни величини. Съвместното разпределение на дискретните величини ξ и η е удобно за намиране на сумата им $\zeta = \xi + \eta$ (случаят на непрекъснати величини е разгледан в §12).

Пример 17.5. (продължение на пример 17.1) Да се намери разпределението на величината $\zeta = \xi + \eta$,

Решение. Сумата на величините, дадени с таблицата от пример 17.1, има възможни стойности са 0, 1 и 2, като

$$P(\zeta=0)=P(\xi=0 \cap \eta=0)=\frac{1}{3}, \quad P(\zeta=2)=P(\xi=1 \cap \eta=1)=\frac{1}{6}$$

$$P(\zeta=1)=P(\xi=0 \cap \eta=1)+P(\xi=1 \cap \eta=0)=\frac{1}{2} \quad (\text{събиране по диагонал})$$

Следователно, за величината ζ получаваме таблицата

ζ	0	1	2
P	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{6}$

♦

Пример 17.6. За да приложим същия метод (чрез събиране на вероятностите, разположени диагонално) за сумата от величините, чието съвместно разпределение е дадено с таблица 17.2,

$\eta \backslash \xi$	-1	0	1
1	0,15	0,3	0,35
3	0,05	0,05	0,1

Табл.17.2.

прибавяме стойност $\eta=2$ с нулеви вероятности (табл.17.3). По този начин сумата от възможните стойности по x_1+y_k , $x_2+y_{k-1}, \dots, x_k+y_1$ е една и съща и затова се събират вероятностите $p_{1,k}+p_{2,k-1}+\dots+p_{k,1}$ ($k=1,2,3$). Така чрез събиране по диагоналите получаваме таблицата

$\xi+\eta$	0	1	2	3	4
P	0,15	0,3	0,4	0,05	0,1

♦

$\eta \backslash \xi$	-1	0	1
1	0,15	0,3	0,35
2	0	0	0
3	0,05	0,05	0,1

Табл.17.3.

Упражнения.

1. Дадено е съвместното разпределение на величините ξ и η :

а)

$x_i \backslash y_j$	1	2	3	4	5	6
0	1/8	1/16	1/16	1/16	1/16	1/8
1	1/12	1/12	1/12	1/12	1/12	1/12

б)

$x_i \backslash y_j$	0	1	2
1/2	5/48	10/48	5/48
1/3	1/27	4/27	4/27
1	0	0	1/4

Да се намерят безусловните разпределения на ξ и η .

2. Имаме една правилна монета, едно правилно зарче и едно зарче, за което вероятността да се паднат 1 или 6 точки е $1/4$, а 2, 3, 4 или 5 точки е $1/8$. Извършва се следният опит: Първо се подхвърля монетата. Ако се падне ези, се хвърля правилното зарче, ако се падне тура, хвърля се неправилното зарче. Да се докаже, че съвместното разпределение на величините ξ - брой на падналите точки и η - брой на падане на ези е дадено с таблицата в задача 1а).

3. В урна има 4 големи и 1 малка червени топки, 2 големи и 3 малки черни топки. Избираме една топка. Величината ξ приема стойност 1, ако изтеглената топка е червена и -1, ако топката е черна. Величината η приема стойност 1, ако изтеглената топка е голяма и -1, ако е малка. Да се намерят:

а) разпределенията на величините ξ , η и (ξ, η) ;

б) разпределенията на величините $\zeta_1 = \xi + \eta$ и $\zeta_2 = \xi \eta$.

4. Двумерната случайна величина е равномерно разпределена в квадрата $D = \{-6 < x < 6; -6 < y < 6\}$, т.е. $p_{\xi,\eta}(x,y)=a$ за $(x,y) \in D$. Да се намери: а) константата a ; б) да се докаже, че ξ и η са независими.

§18. Условни разпределения. Условно математическо очакване. Регресия и линия на регресията.

От съвместния закон на две случайни величини може да намерим безусловните закони на двете съставни величини (но не и обратното). Тези закони не дават информация дали има зависимост между двете

величини. Затова при всяка възможна стойност на едната съставна се съставя закон за разпределение на другата, който се нарича условен.

Изхождайки от формулата за условна вероятност

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

получаваме следните формули за условни вероятности на ξ и η на дискретна случайна величина (ξ, η)

$$P(\xi = x_i | \eta = y_j) = p_{\xi|y_j}(x_i | y_j) = \frac{p_{\xi, \eta}(x_i, y_j)}{p_{\eta}(y_j)}, \quad i=1, \dots, n \quad - \text{закон на}$$

величината ξ , ако $\eta = y_j$.

$$P(\eta = y_j | \xi = x_i) = p_{\eta|x_i}(y_j | x_i) = \frac{p_{\xi, \eta}(x_i, y_j)}{p_{\xi}(x_i)}, \quad j=1, \dots, m \quad - \text{закон на}$$

величината η при $\xi = x_i$.

Ако (ξ, η) е непрекъсната, то условните плътности на компонентите

$$\xi \text{ и } \eta \text{ са } p_{\xi|y}(x|y) = \frac{p_{\xi, \eta}(x, y)}{p_{\eta}(y)}, \quad p_{\eta|x}(y|x) = \frac{p_{\xi, \eta}(x, y)}{p_{\xi}(x)}.$$

За плътността на величината $\xi|\eta=y$ $p_{\xi|y}(x|y)$ x е променлива, а y е параметър, който показва при каква стойност на η е получено условното разпределение на ξ . Аналогично в $p_{\eta|x}(y|x)$ y е променлива, а x е параметър.

Случайните величини ξ и η се наричат независими, ако условният закон на едната величина не зависи от това каква стойност е приела другата, т.е.

$$\begin{aligned} p_{\xi|y_j}(x_i | y_j) &= p_{\xi}(x_i), \quad p_{\eta|x_i}(y_j | x_i) = p_{\eta}(y_j) \quad \text{ако } (\xi, \eta) \text{ е дискретна.} \\ p_{\xi|y}(x|y) &= p_{\xi}(x), \quad p_{\eta|x}(y|x) = p_{\eta}(y), \quad \text{ако } (\xi, \eta) \text{ е непрекъсната} \end{aligned}$$

Пример 18.1. (продължение на пример 17.1). Да се намерят условните закони $\xi|\eta=0$ и $\eta|\xi=1$ и да се определи дали величините са независими.

$$\text{Решение. } p_{\xi|0}(x_1 | \eta=0) = p_{\xi|y_1}(x_1 | y_1) = \frac{p_{\xi, \eta}(x_1, y_1)}{p_{\eta}(y_1)} = \frac{1/3}{2/3} = \frac{1}{2}, \text{ и т.н.}$$

(всички числа от първия ред на съвместното разпределение се разделят на $p_{\eta}(y_1)$). По аналогичен начин пресмятаме и останалите условни вероятности. Така получаваме условните разпределения.

$$\begin{array}{c|cc} x_i | \eta=0 & 0 & 1 \\ \hline p_{\xi|0}(x_i | 0) & \frac{1}{2} & \frac{1}{2} \end{array} \quad \begin{array}{c|cc} y_i | \xi=1 & 0 & 1 \\ \hline p_{\eta|1}(y_i | 1) & \frac{2}{3} & \frac{1}{3} \end{array}.$$

Очевидно, безусловните и условните закони съвпадат, т.е. величините са независими. ♦

Пример 18.2. За величината (ξ, η) със закон на разпределение

η	ξ	-1	0	1
1		0,15	0,3	0,35
2		0,05	0,05	0,1

(виж пример 17.4) да се намери условният закон на ξ при условието $\eta = y_2 = 2$.

$$\text{Решение. Безусловните закони са: } \begin{array}{c|ccc} \xi & -1 & 0 & 1 \\ \hline P & 0,2 & 0,35 & 0,45 \end{array} \quad \begin{array}{c|cc} \eta & 1 & 2 \\ \hline P & 0,8 & 0,2 \end{array}$$

За закона на ξ при условието $\eta = y_2 = 2$ разделяме втори ред на таблицата на безусловната вероятност $p_{\eta}(y_2) = 0,2$ и получаваме

$$\begin{array}{c|ccc} \xi | 2 & -1 & 0 & 1 \\ \hline P & 0,25 & 0,25 & 0,5 \end{array}.$$

Очевидно, ξ и η са зависими. ♦

Пример 18.3. Да се намерят условните закони на величините от пример 17.3.

Решение. Плътността на условното разпределение на величината ξ е $p_{\xi|y}(x|y) = \frac{x+y}{y+0,5}$, където $0 < x < 1$ е променлива, а $0 < y < 1$ е параметър. Например, при $y = 0,5$ плътността на величината $\xi|\eta=0,5$ е $p_{\xi|0,5}(x|0,5) = x + 0,5$.

Аналогично, $p_{\eta|x}(y|x) = \frac{x+y}{x+0,5}$, където $x \in (0,1)$ - параметър, а $y \in (0,1)$ е променлива. ♦

Пример 18.4. Нека съвместното разпределение на две зависими случайни величини се задава с плътността

$$p_{\xi, \eta}(x, y) = \frac{1}{2\pi\sqrt{(1-\rho^2)}} e^{-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}}.$$

а) Да се докаже, че всяка от съставните величини има стандартно нормално разпределение. б) Да се намери условният закон на η при

условие $\xi = x$, където x е произволно число.

Решение. а) Пресмятаме $p_{\xi}(x) = \int_{-\infty}^{\infty} p_{\xi,\eta}(x,y)dy =$

$$\frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} e^{-\frac{x^2-2\rho x+y^2}{2(1-\rho^2)}} dy = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} e^{-\frac{x^2(1-\rho^2)+(x^2\rho^2-2\rho x+y^2)}{2(1-\rho^2)}} dy =$$

$$= \frac{e^{-\frac{x^2(1-\rho^2)}{2(1-\rho^2)}}}{2\pi\sqrt{1-\rho^2}} \sqrt{1-\rho^2} \int_{-\infty}^{\infty} e^{-\frac{(y-\rho x)^2}{2(1-\rho^2)}} d\frac{y-\rho x}{\sqrt{1-\rho^2}} = \frac{e^{-\frac{x^2}{2}}}{2\pi} \sqrt{2\pi} = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}.$$

Тук е използван интегралът на Поасон $\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx = \sqrt{2\pi}$.

По същия начин намираме $p_{\eta}(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$.

б) По формулите за условно разпределение имаме

$$p_{\eta|x}(y|x) = \frac{p_{\xi,\eta}(x,y)}{p_{\xi}(x)} =$$

$$= \frac{\frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{x^2-2\rho x+y^2}{2(1-\rho^2)}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}} = \frac{e^{-\frac{x^2-2\rho x+y^2-x^2+x^2\rho^2}{2(1-\rho^2)}}}{\sqrt{2\pi(1-\rho^2)}} = \frac{1}{\sqrt{2\pi(1-\rho^2)}} e^{-\frac{(y-\rho x)^2}{2(1-\rho^2)}}.$$

следователно, при фиксирано $\xi=x$ величината $\eta|x$ е нормално разпределена с математическо очакване $E(\eta|x)=x\rho$ и средно квадратично отклонение $\sigma=\sqrt{1-\rho^2}$, т.е. със закон $N(x\rho, \sqrt{1-\rho^2})$. Аналогично законът на величината $\xi|y$ е $N(\rho y, \sqrt{1-\rho^2})$. Величините са зависими. ♦

Условно математическо очакване. Често в практиката е необходимо да се оцени характерът на зависимостта между две случайни величини. Например е важно да се намери връзката между количеството валежи и средния добив от единица площ, между параметрите на технологията и качеството на готоввата продукция и т.н. Обикновено едната от двете величини, се счита за дадена, а се търси средната стойност на другата. Както знаем, “носител” на средната стойност на една случайна величина е нейното математическо очакване. Затова, ако $p_{\eta|x}(y|x)$ е плътността на величината $\eta|\xi=x$, то математическото очакване на величината η при зададена стойност $\xi=x$ се пресмята по формулата

$$E(\eta|x) = \int_{-\infty}^{\infty} p_{\eta|x}(y|x)ydy$$

и зависи от стойността x , която величината ξ е приела.

Например, $E(\eta|x)=\rho x$ от пример **19.3** е линейна функция на x .

Аналогично, $E(\xi|y)=\rho y$.

Следователно, ако величините ξ и η са зависими, то $E(\eta|x)$ е функция на x и отразява характера на зависимостта на средната стойност на величината η от стойността, приета от величината ξ .

Функцията $g(x)=E(\eta|x)$ се нарича регресия на η от ξ , функцията $h(y)=E(\xi|y)$ се нарича регресия на ξ от η . Графиките на тези функции се наричат съответно линия на регресия на η от ξ и линия на регресия на ξ от η .

Очевидно, условното математическо очакване е неслучайна функция на случайна величина и затова също се явява случайна величина, означението за която е $E(\eta|x)$. Ще отбележим само свойството

$$E(E(\eta|x))=E(\eta),$$

(математическото очакване на величината $E(\eta|x)$ е равно на математическото очакване на величината η).

Упражнения.

1. За разпределенията от задача 1, §17 да се намерят условните закони на $\eta|\xi=k$, $k=1,2$, и $\xi|\eta=1$ и математическите им очаквания.
2. Дадена е плътността

а) $p_{\xi,\eta}(x,y)=x+y$ за $(x,y)\in D=(0<x<1,0<y<1)$.

б) $p_{\xi,\eta}(x,y)=6x^2y$ за $(x,y)\in D=(0<x<1,0<y<1)$

Да се намерят регресиите на η от ξ и на ξ от η и да се начертаят графиките им.

§19. Ковариация. Коефициент на корелация.

Ковариация $\text{cov}(\xi,\eta)$ на две случайни величини ξ и η се нарича математическото очакване на величината $(\xi-E\xi)(\eta-E\eta)$, т.е.

$$\text{cov}(\xi,\eta)=E[(\xi-E\xi)(\eta-E\eta)]. \quad (19.1)$$

Ако $\text{cov}(\xi,\eta)=0$, то величините ξ и η се наричат неколерирани.

Ако величините ξ и η са независими, то

$$\text{cov}(\xi,\eta)=E[(\xi-E\xi)(\eta-E\eta)]=E[(\xi-E\xi)]E[(\eta-E\eta)]=(E\xi-E\xi)(E\eta-E\eta)=0$$

Следователно, ако величините са независими, то те са неколерирани. Обратното, обаче, не винаги е вярно, т.е. може $\text{cov}(\xi,\eta)=0$, но величините да са зависими.

Ако $\text{cov}(\xi, \eta) \neq 0$, то величините ξ и η се наричат **корелирани**.

Пример 19.1. Ще намерим ковариацията на двумерна случайна величина със съвместното разпределение

x_i	-1	0	1
y_j	-1	0	1
-1	0	0,25	0
0	0,25	0	0,25
1	0	0,25	0

Решение.

Тъй като $E\xi = E\eta = 0$. То $\text{cov}(\xi, \eta) = E[(\xi - 0)(\eta - 0)] = E(\xi\eta)$.

За всички двойки (x_i, y_j) или x_i , или y_j , или на p_{ij} е равно на нула, откъдето

$$E(\xi\eta) = 0 \text{ и } \text{cov}(\xi, \eta) = E(\xi\eta) - E\xi E\eta = 0,$$

т.е. величините са некорелирани. Въпреки това те са зависими - ако едната е приела стойност, различна от нула, то другата трябва да бъде равна нула. ♦

За изчисляване на ковариацията вместо по формула (19.1) се използва формулата

$$\text{cov}(\xi, \eta) = E(\xi\eta) - E\xi E\eta, \quad (19.2)$$

където $E\xi$ и $E\eta$ са математическите очаквания на величините ξ и η

$$E(\xi\eta) = \begin{cases} \sum_{i=1}^n \sum_{j=1}^m p_{\xi, \eta}(x_i, y_j) \cdot x_i \cdot y_j & (\xi, \eta) - \text{дискретна} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \cdot y \cdot p_{\xi, \eta}(x, y) dx dy & (\xi, \eta) - \text{непрекъсната} \end{cases}$$

Доказателство. Тъй като $E\xi$ и $E\eta$ са константи, то

$$\text{cov}(\xi, \eta) = E(\xi\eta - \xi \cdot E\eta - \eta \cdot E\xi + E\xi E\eta) = E(\xi\eta) - E\xi \cdot E\eta - E\eta \cdot E\xi + E\xi \cdot E\eta,$$

Откъдето следва формула (19.2). ♦

Други свойства на ковариацията са:

$$(\text{cov}(\xi, \eta))^2 \leq D\xi D\eta$$

Доказателство. Да използваме, че за дисперсията на случайната величина $a\xi + b\eta$, където a и b са произволни реални числа, е изпълнено неравенството $D(a\xi + b\eta) \geq 0$. От свойствата на дисперсията имаме

$$\begin{aligned} D(a\xi + b\eta) &= E(a\xi + b\eta - (aE\xi + bE\eta))^2 = E(a(\xi - E\xi) + b(\eta - E\eta))^2 = \\ &= a^2 E(\xi - E\xi)^2 + 2abE((\xi - E\xi)(\eta - E\eta)) + b^2 E(\eta - E\eta)^2 = \\ &= a^2 D\xi + b^2 D\eta + 2abE((\xi - E\xi)(\eta - E\eta)) \end{aligned}$$

Следователно,

$$D(a\xi + b\eta) = a^2 D\xi + b^2 D\eta + 2ab \text{cov}(\xi, \eta). \quad (19.3)$$

Като разделим на b^2 и означим $t = \left(\frac{a}{b}\right)$, получаваме квадратното

неравенство $D\xi \cdot t^2 + 2\text{cov}(\xi, \eta) \cdot t + D\eta \geq 0$. То е изпълнено за всяко t , ако дискриминантата му е неотрицателна, т.е. $\text{cov}^2(\xi, \eta) - D\xi D\eta \geq 0$. ♦

Следствие 19.1. От формула (19.3) при $a=1$, $b=1$ получаваме

формула за дисперсия на сума на две зависими величини

$$D(\xi + \eta) = D\xi + D\eta + 2\text{cov}(\xi, \eta)$$

Като характеристика на степента на зависимост на случайните величини се използва

$$\text{коэффициентът на корелация } \rho(\xi, \eta) = \frac{\text{cov}(\xi, \eta)}{\sqrt{D\xi D\eta}}.$$

Свойства на коефициента на корелация:

- 1) $|\rho(\xi, \eta)| \leq 1$.
- 2) $\rho(\xi, \eta)$ е безразмерна величина.
- 3) ако ξ и η са независими, то $\rho(\xi, \eta) = 0$.
- 4) ако $\eta = a\xi + b$, т.е. ξ и η са линейно зависими, то $\rho(\xi, \eta) = 1$.

В общия случай, ако $\rho(\xi, \eta) = 0$, не може да се твърди, че ξ и η са независими. За некорелирани случайни величини $\rho(\xi, \eta) \neq 0$,

Пример 19.2. Ще намерим ковариацията и коефициента на корелация на величината от пример 17.4.

$$E\xi\eta = 0,15 \cdot 1 \cdot (-1) + 0,3 \cdot 0,1 + 0,35 \cdot 1,1 + 0,05 \cdot 1,2 + 0,05 \cdot 0,1 + 0,1 \cdot 1,1 = 0,4,$$

$$\text{cov}(\xi, \eta) = E\xi\eta - E\xi E\eta = 0,4 - 0,5875 \cdot 0,16 = 0,306,$$

$$\rho(\xi, \eta) = \frac{\text{cov}(\xi, \eta)}{\sqrt{D\xi D\eta}} = \frac{0,1}{\sqrt{0,5875 \cdot 0,16}} = 0,326. \quad \spadesuit$$

Упражнения.

1. Дадена е $\text{cov}(\xi, \eta) = 3$. Да се намери $\text{cov}(2\xi - 5, 4\eta + 2)$.
2. Да се намерят ковариацията и коефициента на корелация на величините, за които а) $p_{\xi, \eta}(x, y) = x + y$ за $(x, y) \in D = (0 < x < 1, 0 < y < 1)$.
б) $p_{\xi, \eta}(x, y) = 6x^2 y$ за $(x, y) \in D = (0 < x < 1, 0 < y < 1)$
в) $p_{\xi, \eta}(x, y) = 15x^2 y$ за $(x, y) \in D = (0 < x < y < 1)$
3. Да се намерят ковариацията и коефициента на корелация на величината (ξ, η) от задача 3, §17.
4. Да се намерят математическите очаквания и дисперсиите на величините ζ_1 и ζ_2 от задача 3, §17, и проверят свойствата им.

СТАТИСТИКА

§20. Предмет и задачи на математическата статистика. Генерална съвкупност и извадка.

Математическата статистика е приложен дял от математиката, основа на която е теорията на вероятностите. Тя е *наука за събиране, обработване и интерпретиране на числови данни*. Основната ѝ цел е с помощта на събраните данни да се даде отговор на реални въпроси като:

- Има ли изменение в климата;
 - Какво е въздействието на новосъздадено лекарство;
 - Какви са тенденциите в икономиката на дадена страна;
 - Кой от два метода на обучение е по-ефективен;
- т.е. въпроси, свързани с явления и обекти, които или не са изучени, или е невъзможно пълното им изследване. За да бъде намерен правилният им отговор са създадени
- методи за събиране на данни. Тук основен въпрос е с минимални разходи да се възможно най-достоверна съвкупност от данни.
 - методи за обработка и интерпретация на данните. С помощта на тези методи се правят изводи за развитието и закономерностите на явленията.

Основен предмет на математическата статистика е изследването на съвкупности от обекти, която не можем да изучим напълно.

Някои основни понятия:

Генерална съвкупност (популация) - съвкупността от еднотипни обекти, която се изследва. Обикновено се изследва един или няколко характеристики (признаци) на генералната съвкупност, които получават различни стойности преминавайки от един обект към друг обект на съвкупността. Тези характеристики ще наричаме променливи и ще ги означаваме с главни букви X , Y , Z и т.н.

Променливите (признаците) могат да бъдат:

качествени, ако стойността им не може да се представи като число.
количествени, ако стойността им има смисъл на число (измерване или броене).

Например, изследват се жителите в даден град (генерална съвкупност) по ръст (количествен признак) и по цвят на очите (качествен признак).

Когато анализираме качествен признак (например изучаваме структурата на икономиката) обикновено се работи с брой или проценти – какъв процент от общата продукция се пада на селското стопанство, на промишлеността, на транспорта и т.н. В най-простия случай, качественият признак се свежда към количествен като се въведе величина с две възможни стойности $X=1$, ако даден обект от генералната съвкупност притежава качеството, в противен случай $X=0$. В по-сложните случаи на определена стойност на качественият признак може да се присвои числена стойност, но действията между такива

числени стойности нямат никакъв реален смисъл. Например, в преброяването на населението са въведени следните категории: несемеен, семеен, разведен и вдовец, на които може да се припишат числени стойности, примерно 0, 1, 2 и 3 (тук сумата $2+1$ не изразява нищо). Ще отбележим, че със същия успех може да се вземат и произволни други четири символа, например, ♣, ♦, ♥, и ♠.

Количествените променливи (признаци) се делят на две групи:

Дискретни, ако множеството от всички възможни стойности е дискретно.

Непрекъснати, ако множеството от всички възможни стойности се състои от интервал.

Вниманието ни ще бъде насочено към количествените променливи като в зависимост от интересите ни, се разглеждат една, две или повече количествени характеристики на генералната съвкупност – едномерен, двумерен и многомерен признак на генерална съвкупност.

Изследването на генералната съвкупност може да протече по следните начини:

- Да се изследват всички обекти от генералната съвкупност. По този начин се описват съвсем точно характеристиките на генералната съвкупност. Този метод е много неефективен, а в някои случаи и невъзможен (например, ако изследването води до разрушаване на обекта).

- Да се изследва само част от генералната съвкупност (извадка). Предимство тук е спестяването на средства и време. Недостатък е, че не се получава абсолютно точно описание на генералната съвкупност. Затова е от голямо значение подборът на обектите за изследване.

Пример 20.1. Изследва се средният ръст на населението на дадена страна. Това може да стане, ако вместо цялото население се изберат отделни индивиди и да се изчисли средният ръст за избраната група. Очевидно, много важно е как ще се избират отделните индивиди - необходимо е групата да бъде достатъчно голяма, индивидите да са избрани от различни области и случайно (а не избирателно, например, които са по-високи). ♦

Извадка наричаме частта от генералната съвкупност, която е достъпна за изследване. Броят на елементите в извадката се нарича обем на извадката.

Репрезентативна (представителна) извадка е тази, която отразява достоверно генералната съвкупност. Ако изборът на елементите става по случаен начин, то извадката се нарича случайна.

От опит е установено, че репрезентативни са случайните извадки с голям обем. Затова ще разглеждаме само случайни извадки. Тъй като изборът на елементи от генералната съвкупност е случаен, то променливата която се изучава, ще разглеждаме като случайна величина X с неизвестен закон, а резултатите от всяко измерване (или броене) –

възможни стойности $x^{(1)}, x^{(2)}, \dots, x^{(n)}$ на случайни величини $X_1, X_2, \dots, X_n$ със същия закон на разпределение. По този начин:

При изследване на количествената променлива X на генералната съвкупност:

- величината X се счита за случайна величина с неизвестен закон;
- извадката се представя като многомерна случайна величина $(X_1, \dots, X_n)$, чиито компоненти $X_i = \{\text{резултат от } i\text{-тото наблюдение}\}$ имат разпределение, съпадащо с разпределението на X ;
- наблюдаваните стойности $x^{(1)}, x^{(2)}, \dots, x^{(n)}$ са реализация на величините $X_1, X_2, \dots, X_n$, т.е. наблюдавани са събитията $X_1 = x^{(1)}, X_2 = x^{(2)}$ и т.н.

Следователно, теоретична основа на статистиката е теорията на вероятностите, с помощта на която се решават следните задачи:

- Основни задачи на математическата статистика:**
1. Да се систематизират събраните данни (статистическа обработка на данните).
 2. Въз основа на данните да се оцени неизвестна количествена характеристика на съвкупността (оценка на параметри).
 3. Да се направят изводи, отнасящи се за цялата съвкупност и провери верността им (проверка на хипотези).
 4. Ако се изследват зависимости между две или повече променливи, да се установи характера и степента им на зависимост една от друга (корелационен или регресионен анализ).

СТАТИСТИЧЕСКА ОБРАБОТКА НА ДАННИ

§21. Статистическо разпределение, полигон и хистограма.

Емпирична функция на разпределение.

Статистиката започва със систематизиране на събраните данни. Първата стъпка е данните да се представят в такава форма, че да може да се състави обща картина на изследванията. За да се интерпретира по-нататък тази съвкупност от данни, като се има предвид източника на данните и методите на събирането им, трябва да се отстранят тези от данните (ако има такива), които са попаднали случайно или по погрешка в изследваната съвкупност. До голяма степен това се прави по осматриване на изследователя. Целта ни тук е да представим данните в такъв вид, че да може да се вземе решение по този въпрос и да се даде насока за по-нататъчните изследвания.

Нека изследваме количествения признак X на генералната съвкупност, за която разполагаме с резултатите от n наблюдения, т.е. дадена ни е извадка $(X_1, X_2, \dots, X_n)$ с обем n . Тъй като в резултат на тези

наблюдения получаваме в n стойности $(x^{(1)}, x^{(2)}, \dots, x^{(n)})$ на променливата X , то тези стойности, се наричат варианти. Подреждаме вариантите по големина (между тях може да има и повторения). Така получаваме

Вариационен ред - наблюдаваните стойности, наредени по големина

$$\hat{x}_1 \leq \hat{x}_2 \leq \dots \leq \hat{x}_n. \quad (21.1)$$

Първите оценки на съвкупността от данни са

Размах и среда на извадката: $x_{\max} - x_{\min}$ и $\frac{x_{\max} + x_{\min}}{2}$.

1. Статистическо разпределение. Ако нанесем в таблица само различните варианти и броят на повторенията им (честота), получаваме

Статистическо разпределение на честотите - таблица от вида

$$\begin{array}{c|cccc} X & x_1 & x_2 & \dots & x_k \\ \hline m_i & m_1 & m_2 & \dots & m_k \end{array}, \quad \sum_{i=1}^k m_i = n, \quad (21.2)$$

където m_i е честота на вариантата x_i .

Отношението $v_i = \frac{m_i}{n}$ наричаме относителна честота на вариантата x_i . Таблицата

$$\begin{array}{c|cccc} X & x_1 & x_2 & \dots & x_k \\ \hline v_i & v_1 & v_2 & \dots & v_k \end{array}, \quad \sum_{i=1}^k v_i = 1 \quad (21.3)$$

наричаме статистическо разпределение на относителните честоти.

Графически статистическите разпределения се представят с

Полигон на честотите: начупена линия, съединяваща точките с координати (x_i, m_i) или полигон на относителните честоти: - начупена линия, съединяваща точките с координати (x_i, v_i) (фиг. 21.1).

Очевидна е връзката между разпределението на една дискретна случайна величина и статистическото разпределение на относителните честоти. По аналогия се въвежда и функцията

Емпирична функция на разпределението:

$$F^*(x) = \frac{n_x}{n}, \quad x \in (-\infty, \infty), \quad \text{където } n_x - \text{брой на вариантите, по-малки от } x.$$

Като се използват:

$$\text{натрупаните (кумулятивни) честототи } c_i = \sum_{j=1}^i m_j \text{ и}$$

относителните натрупани (кумулятивни) честоти $\gamma_i = \frac{c_i}{n} = \sum_{j=1}^i \nu_j$

не е трудно да се види, че

$$F^*(x) = \begin{cases} 0 & x \leq x_1 \\ \gamma_i & x_1 < x \leq x_{i+1}, i=1, \dots, k-1, \\ 1 & x > x_k \end{cases} \quad (21.4)$$

Забележка 21.1. Както знаем, функцията на разпределение на една случайна величина се използва преди всичко за намиране на квантили на случайната величина. Използването на емпиричната функция на разпределение е аналогично, но за да може да се използва с успех във всички случаи (виж §23) е по-удобно въвеждането на функция, която да бъде непрекъсната за всяко x . Тази функция ще наричаме

кумулятивна функция

$$F(x) = \begin{cases} 0 & x \leq x_1 \\ \gamma_2 + \frac{\gamma_2}{x_2 - x_1}(x - x_1) & x_1 < x \leq x_2 \\ \gamma_i + \frac{\nu_i}{x_i - x_{i-1}}(x - x_{i-1}) & x_{i-1} < x \leq x_i, i=3, \dots, k \\ 1 & x > x_k \end{cases} \quad (21.5)$$

Графиката на емпиричната функция на разпределението $F^*(x)$ е стъпаловидна, а графиката на кумулативната функция $F(x)$, която се нарича също полигон на относителните кумулативни честоти се състои от отсечки, съединяващи точките с координати $(x_1, 0)$, (x_i, γ_i) ($i=2, \dots, k$).

Забележка 21.2. Единствената разлика между вариационния ред (21.1) и статистическото разпределение (21.2) е в това, че данните в (21.2) са групирани по стойност. В частност, сумата S на всички наблюдавани стойности, може да намерим по два начина: ако използваме вариационния ред то $S = \hat{x}_1 + \hat{x}_2 + \dots + \hat{x}_n$, ако използваме

статистическото разпределение. то $S = m_1 x_1 + m_2 x_2 + \dots + m_k x_k = \sum_{i=1}^k m_i x_i$,

Пример 21.1. Броят на аварияте в едно предприятие през последните десет години е 5,2,3,4,5,4,0,3,4,3. Да се намери размахът и средата на извадката, статистическото разпределение на честотите и натрупаните честоти, емпиричната и кумулативната функция. Да се

начертаят полигонът на относителните честоти и графиката на емпиричната функция на разпределението.

Решение. Очевидно, размахът на извадката е $5 - 0 = 5$,

средата е $\frac{5+0}{2} = 2,5$.

Всички необходими изчисления нанасяме в таблица:

i	варианти x_i	честоти m_i	относит. честоти $\nu_i = \frac{m_i}{n}$	относит. кумулят. честоти. $\gamma_i = \frac{c_i}{n}$	$\frac{\nu_i}{x_i - x_{i-1}}$
1	0	1	0,1	0,1	—
2	2	1	0,1	0,2	0,05
3	3	3	0,3	0,5	0,3
4	4	3	0,3	0,8	0,3
5	5	2	0,2	1,0	0,2

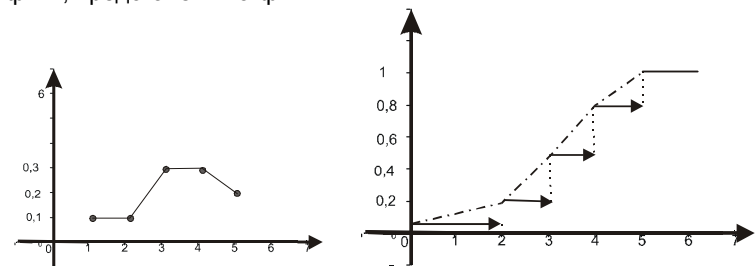
Статистическото разпределение на относителните честоти е

$$\begin{array}{c|ccccc} x_i & 0 & 2 & 3 & 4 & 5 \\ \hline \nu_i & 0,1 & 0,1 & 0,3 & 0,3 & 0,2 \end{array}, \quad (\nu_i = \frac{m_i}{n} = \frac{m_i}{10})$$

а полигонът на относителните честоти е начертан на фиг.21.1. По формули (21.3) и (21.7) получаваме емпиричната функция на разпределението $F^*(x)$ и кумулативната функция $F(x)$

$$F^*(x) = \begin{cases} 0 & x < 0 \\ 0,1 & 0 \leq x < 2 \\ 0,2 & 2 \leq x < 3 \\ 0,5 & 3 \leq x < 4 \\ 0,8 & 4 \leq x < 5 \\ 1 & x \geq 5 \end{cases} \text{ и } F(x) = \begin{cases} 0 & x < 0 \\ 0,2 + 0,05(x-2) & 0 \leq x < 2 \\ 0,5 + 0,3(x-3) & 2 \leq x < 3 \\ 0,8 + 0,3(x-4) & 3 \leq x < 4 \\ 1 + 0,2(x-5) & 4 \leq x < 5 \\ 1 & x \geq 5 \end{cases}$$

с графики, представени на фиг. 21.1.



Фиг. 21.1

Фиг.21.2

Забележка 21.3. В интервала $(0,2]$ функцията $F(x)$ е построена по правилото за останалите интервали, затова има прекъсване за $x=0$. За

да се получи функцията (21.5), числото във втория ред на последната колона в таблицата трябва да бъде $\gamma_2/(x_2-x_1)=0,2:2=0,1$. ♦

2. Интервално статистическо разпределение.

Ако X е непрекъсната случайна величина или обемът на извадката е много голям, е целесъобразно да се постъпи по следния начин. Разделяме интервала $[x_{\min}, x_{\max}]$ на k подинтервали $[x_0, x_1), \dots, [x_{k-1}, x_k)$ (препоръчително е броят на подинтервалите да е между 5 и 15). Означаваме с m_i броя на вариантите, попадащи $[x_{i-1}, x_i)$. Така получаваме

$$\begin{array}{c|c|c|c|c|c} [x_{i-1}, x_i) & [x_0, x_1) & [x_1, x_2) & \dots & [x_{k-1}, x_k) & \\ \hline m_i & m_1 & m_2 & \dots & m_k & \end{array}, \quad \sum_{i=1}^k m_i = n, \quad (21.6)$$

интервалното статистическо разпределение на относителните честоти:

$$\begin{array}{c|c|c|c|c|c} [x_{i-1}, x_i) & [x_0, x_1) & [x_1, x_2) & \dots & [x_{k-1}, x_k) & \\ \hline v_i & v_1 & v_2 & \dots & v_k & \end{array}, \quad v_i = \frac{m_i}{n}, \quad \sum_{i=1}^k v_i = 1 \quad (21.7)$$

Използват се също и

$$\text{натрупана (кумулятивна) честота } c_i = \sum_{j=1}^i m_j;$$

$$\text{относителна натрупана (кумулятивна) честота } \gamma_i = \frac{c_i}{n} = \sum_{j=1}^i v_j.$$

Разпределенията (21.6) и (21.7) се представят чрез графики които се наричат хистограми (фиг.21.3).

Хистограма на честотите: фигура, състояща се от правоъгълници, с основа $[x_{i-1}, x_i)$ и лице, равно на честотата m_i , т.е. с височина

$$h_i = \frac{m_i}{x_i - x_{i-1}}.$$

Хистограма на относителните честоти: фигура, състояща се от правоъгълници, с основа $[x_{i-1}, x_i)$ и лице, равно на v_i .

За по-нататъчната обработка на интервалното разпределение, с цел да се получи разпределение от вида (21.2) или (21.3), интервалът $[x_{i-1}, x_i)$ се заменя със средата му $x_i^* = \frac{x_{i-1} + x_i}{2}$ ($i=1, \dots, k$) и така се получава разпределението

$$\begin{array}{c|c|c|c|c} x_1^* & x_2^* & x_2^* & \dots & x_k^* \\ \hline m_i & m_1 & m_2 & \dots & m_k \end{array}. \quad (21.8)$$

Сега можем да получим и други характеристики на извадката, които ще бъдат разгледани по-нататък. Например, по формула (21.3) получаваме емпиричната функция $F^*(x)$ на разпределението, но по-удобна за приложение е следната непрекъсната функция, наречена

Кумулативна функция:

$$F(x) = \begin{cases} 0 & x \leq x_0 \\ \gamma_i + \frac{v_i}{d_i}(x - x_i) & x_{i-1} < x \leq x_i, \quad i=1, \dots, k. \\ 1 & x > x_k \end{cases} \quad (21.9)$$

където $d_i = x_i - x_{i-1}$ е дължината на интервала $[x_{i-1}, x_i)$

Графиката на функцията $F(x)$ в интервала $(x_0, x_k]$ се състои от отсечки, съединяващи точките $(x_0, 0), (x_1, \gamma_1), \dots, (x_k, \gamma_k)$.

Забележка 21.4. Кумулативната функция както и графиката ѝ са много удобни при определяне на разгледаните в следващия параграф квантили.

Пример. 21.2. Дадена е извадка с обем $n=20$: 4,5; 1,2; 3,0; 4,0; 1,1; 4,2; 2,1; 3,4; 2,4; 3,9; 3,7; 4,3; 4,9; 2,7; 2,7; 3,6; 3,3; 4,2; 3,9; 2,4. Да се състави интервален статистически ред и построи хистограмата на честотите. Да се намерят емпиричната и кумулативната функции на разпределение и построят графиките им.

Решение: Подреждаме числата в групи по цялата им част:

1,2; 1,1;
2,1; 2,4; 2,7; 2,7; 2,4;
3,0; 3,4; 3,9; 3,7; 3,6; 3,3; 3,9.
4,5; 4,0; 4,2; 4,3; 4,9; 4,2,

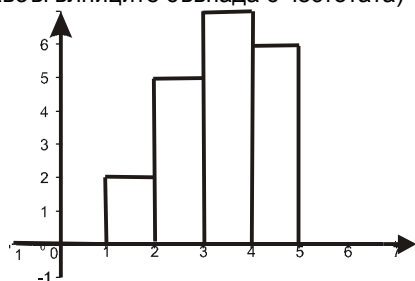
откъдето получаваме интервалното статистическо разпределение

$$\begin{array}{c|c|c|c|c} [x_i, x_{i+1}) & [1, 2) & [2, 3) & [3, 4) & [4, 5) \\ \hline m_i & 2 & 5 & 7 & 6 \end{array}.$$

По-нататък ще използваме следната таблица, в която са отразени всички необходими пресмятания:

i	интервал $[x_i, x_{i+1})$	честота m_i	кумулят. честота $c_i = \sum_{j=1}^i m_j$	относит. честота $v_i = \frac{m_i}{n}$	относит. кумулят. честота $\gamma_i = \frac{c_i}{n}$	среда на интервала $x_i^* = \frac{x_i + x_{i-1}}{2}$
1	[1, 2)	2	2	0,10	0,10	1,5
2	[2, 3)	5	7	0,25	0,35	2,5
3	[3, 4)	7	14	0,35	0,70	3,5
4	[4, 5)	6	20	0,30	1,00	4,5

Хистограмата на честотите начертаваме като използваме втората колона (понеже дължината на интервалите $[x_{i-1}, x_i)$ е единица, височината на правоъгълниците съвпада с честотата)

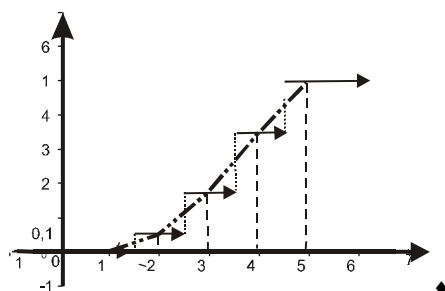


Фиг.21.3.

От тук, съгласно формули (21.8) и (21.9) получаваме емпиричната и кумулативната функции:

$$F^*(x) = \begin{cases} 0 & x < 0,5 \\ 0,1 & 0,5 \leq x < 2,5 \\ 0,35 & 2,5 \leq x < 3,5 \\ 0,7 & 3,5 \leq x < 4,5 \\ 1 & x \geq 4,5 \end{cases}, \quad F(x) = \begin{cases} 0 & x < 1 \\ 0,1 + 0,1(x-1) & 1 \leq x < 2 \\ 0,35 + 0,25(x-2) & 2 \leq x < 3 \\ 0,7 + 0,35(x-3) & 3 \leq x < 4 \\ 1 + 0,3(x-4) & 4 \leq x < 5 \\ 1 & x \geq 5 \end{cases}$$

а техните графики са начертани на фиг. 21.4.



Фиг. 21.4.

Упражнения

1. Извадката 5, 3, 7, 10, 5, 5, 2, 10, 7, 2, 7, 7, 4, 2, 4 представете като а) вариационен ред; б) неинтервално статистическо разпределение. Да се намери емпиричната функция на разпределение; в) интервално статистическо разпределение с интервали (1,4], (4,8], (8,12]. Да се намери кумулативната функция на разпределение.

2. Да се построи хистограмата на честотите на извадката

(x_{j-1}, x_j)	(10,12)	(12,14)	(14,16)	(16,18)	(18,20)	(20,22)	(22,24)
m_j	2	4	8	12	16	10	3

Да се намерят кумулативната и емпиричната функция и да се начертаят графиките им.

3. Да се намери емпиричната функция на разпределение на извадката, дадена със статистическия ред $\frac{x_i}{m_i} \begin{matrix} 2 & 5 & 7 & 8 \\ 1 & 3 & 2 & 4 \end{matrix}$.

4. Записан е броят на грешките върху една страница от текст:

2, 3, 6, 1, 4, 1, 2, 0, 1, 5
1, 4, 4, 2, 3, 1, 2, 2, 1, 0.

Да се начертае полигонът на относителните честоти и намери емпиричната функция.

5. Да се намери интервално статистическо разпределение на извадката (брой на пътниците в един автобус по дадено направление):

16 35 29 30 37 20 34 40 46 10 23 48 33 45
50 32 12 15 13 25 25 34 10 14 39 20 25.

Да се построи хистограма на относителните честоти.

§22. Числени характеристики на статистическото разпределение.

Характеристики на местоположението - мода, медиана, квантили и квартили.

Нека се изследва количествената променлива X на генералната съвкупност и нека е дадена извадка $(X_1, X_2, \dots, X_n)$ с обем n . Освен графически, изследването на извадката се извършва с помощта на числени характеристики – мерки (числа), които дават допълнителна информация за извадката.

Числените характеристики на извадката се разделят на:

- характеристики на местоположението;
- характеристики на разсейването.

Например, размахът и средата са характеристики съответно на разсейването и на местоположението. Освен тези първи и недостатъчно прецизни мерки, от гледна точка на местоположението на вариантите y са въведени следните характеристики на извадката:

Мода - наблюдаваната стойност M_o с най-голяма честота.

Една извадка може да има повече от една мода и се нарича бимодална, ако има две моди и полимодална, ако модите са повече от две. Ако извадката е зададена с интервално разпределение, то интервалът с най-голяма честота се нарича **модален интервал**.

Медиана – такова число Q_2 , за което половината от вариантите са по-малки, а другата половина – по големи от него. (друго означение M_e).

Квантил от ред p ($0 < p < 1$) – число x_p , за което $100p\%$ от вариантите са по-малки от него.

Очевидно, $Q_2 = x_{0,5}$.

Заедно с медианата ($p=0,5$), особено често се използват квантилите от ред $p=0,25$ и $p=0,75$, които носят имената:

Долен квантил - $Q_1 = x_{0,25}$ - горен квантил - $Q_3 = x_{0,75}$.

Друга характеристика, които се изразяват чрез квантилите са:

процентил $x_{p\%}$ от ред $P\%$ ($0 < P < 100$) – число, съвпадащо с квантил от ред $p = P/100$.

Всяка числена характеристика се пресмята на базата на наблюдаваните данни ($x^{(1)}, x^{(2)}, \dots, x^{(n)}$), затова се явява тяхна функция.

Формулите за пресмятане на тези величини зависят от това дали е дадено статистическо или интервално статистическо разпределение на извадката.

1. Статистическо разпределение (неинтервално), за определяне на квантилите е по-удобно е да се използва вариационният ред $\hat{x}_1 \leq \hat{x}_2 \leq \dots \leq \hat{x}_n$ (който винаги може да получим от разпределението)

Квантилът x_p (от ред p) се изчислява по следния начин:

1) Означаваме с r цялата част на числото np . Нека $\hat{x}_r$ е тази варианта, за която np наблюдения са по-малки или равни на нея).

2) Квантилът от ред p е числото

$$x_p = \begin{cases} \frac{\hat{x}_r + \hat{x}_{r+1}}{2}, & \text{ако } np \text{ е цяло} \\ \hat{x}_{r+1}, & \text{ако } np \text{ не е цяло} \end{cases} \quad (22.1)$$

За $p = \frac{1}{2}$ се получава формулата за медианата:

$$Q_2 = \begin{cases} \frac{\hat{x}_r + \hat{x}_{r+1}}{2}, & \text{ако } n \text{ четно} \\ \hat{x}_{r+1}, & \text{ако } n \text{ нечетно} \end{cases}, \text{ където } r \text{ цялата част на числото } \frac{n}{2}$$

(в ляво от $\hat{x}_r$ имаме не повече от половината наблюдения).

Пример 22.1. Да се определи на медианата:

а) За вариационния ред 3, 5, 6, 7, 7. Имаме: $n=5$ - нечетно, цялата част на $\frac{n}{2}=2,5$ е 2. Следователно $r=2$ и $Q_2 = \hat{x}_{2+1} = \hat{x}_3 = 6$.

б) За вариационния ред 3, 5, 6, 7, 7, 9. Тук $Q_2 = \frac{\hat{x}_3 + \hat{x}_4}{2} = \frac{6+7}{2}$ ($r = \frac{n}{2} = \frac{6}{2} = 3$).

в) За вариационния ред 3, 5, 5, 6, 6, 7, 7 $Q_2 = 6$. ♦

Пример 22.2: За вариационния ред (с обем на извадката $n=20$)

0, 0, 0, 1, 2, 2, 3, 3, 6, 6, 6, 7, 7, 8, 8, 12, 12, 20, 20, 20

да се намерят:

$$x_{0,2} = \frac{\hat{x}_4 + \hat{x}_5}{2} = \frac{1+2}{2} = 2,5, \text{ тъй като } np = 20 \cdot 0,2 = 4 \text{ е цяло число, а}$$

$$x_{\frac{2}{3}} = \hat{x}_{13+1} = 8, \text{ защото } np = 20 \cdot \frac{2}{3} = 13,3333 \text{ не е цяло и } r=13.$$

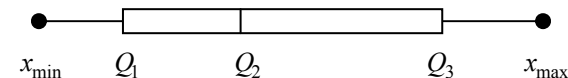
$$x_{40\%} = x_{0,4} = \frac{\hat{x}_8 + \hat{x}_9}{2} = \frac{3+6}{2} = 4,5 \text{ (} np = 20 \cdot 0,4 = 8 \text{ - цяло).}$$

$$Q_1 = \frac{\hat{x}_5 + \hat{x}_6}{2} = \frac{2+2}{2} = 2, Q_2 = \frac{\hat{x}_{10} + \hat{x}_{11}}{2} = \frac{6+6}{2} = 6, Q_3 = \frac{\hat{x}_{15} + \hat{x}_{16}}{2} = \frac{8+12}{2} = 10. ♦$$

Забележка 22.1. Горните формули не са много точни, особено, когато има и други варианти, равни на x_r (какъвто е случаят за Q_1 и Q_2) или обемът n на извадката е твърде малък. По-добре е да се използва кумулативната функция (21.5) като не е необходимо пълното ѝ определяне, а само в интервала, в който се намира квантилът (пример 22.5).

Квантилите Q_1, Q_2, Q_3 , най-малката $x_{\min} = \hat{x}_1$ и най-голямата $x_{\max} = \hat{x}_n$ наблюдавана стойност се използват за да се даде едно просто и нагледно представяне на данните – така наречената

диаграма от тип “кутия” (box-plot) - диаграма от вида



Тази диаграма представя най-важните елементи на извадката: половината от данните се намират в интервала (Q_1, Q_3) , т.е. те са “вътре в кутията”, а медианата Q_2 като число, за което половината от данните са по-големи, а другата половина – по-малки, дава информация за симетричността на данните.

Понякога в извадката попадат данни, които не са типични за генералната съвкупност – варианти, които в сравнение с останалите са или много малки, или много големи. Едно правило за определяне на нетипичните данни е следното:

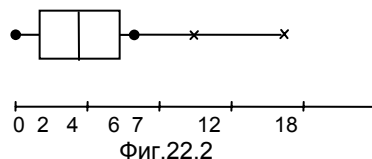
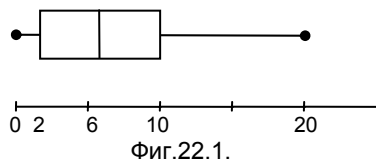
Като нетипични се приемат вариантите, които са:

по-малки от числото $l_1 = Q_1 - 1,5(Q_3 - Q_1)$

по-големи от числото $l_2 = Q_3 + 1,5(Q_3 - Q_1)$.

Пример 22.3. Да се начертае диаграмата от тип кутия на извадката от пример 22.2.

Решение. Имаме: $x_{\min}=0$, $x_{\max}=20$, $Q_1=2$, $Q_2=6$, $Q_3=10$. Диаграмата е начертана на фиг. 22.1.



Пример 22.4. Получена е следната извадка (брой на закъснелите за работа): 0, 1, 2, 2, 3, 4, 5, 5, 6, 7, 18. Да се определи кои от данните са нетипични и начертае диаграмата от тип кутия.

Решение: Пресмятаме $Q_1=\hat{x}_3=2$, $Q_2=x_6=4$, $Q_3=x_9=6$,
 $I_1=2-1,5(6-2)=-4$, $I_2=6+1,5(6-2)=12$.

Следователно, нетипична е вариантата $\hat{x}_{11}=18$, защото е по-голяма от I_2 . На диаграмата тази стойност е отбелязана с кръстче (фиг.22.2). ♦

2. Интервално статистическо разпределение. Нека е дадена таблицата

$[x_{i-1}, x_i]$	$[x_0, x_1]$	$[x_1, x_2]$	...	$[x_{k-1}, x_k]$
m_i	m_1	m_2	...	m_k

За намирането квантил от ред p е добре да са изчислени относителните кумулативни честоти $\gamma_i = \sum_{j=1}^i v_j$, където $v_j = \frac{m_j}{n}$. Тогава интервалът, в който се намира квантилът е първият интервал $[x_{r-1}, x_r]$, за който $\gamma_r \geq p$. Като се използва формула (22.8) постъпваме по следния начин:

Намиране на квантил x_p от ред p :

- 1) Намираме най-малката относителна кумулативна честота γ_r , за която $p \leq \gamma_r$
- 2) Интервалът, в който се намира квантилът, е $[x_{r-1}, x_r]$.
- 3) Намираме квантила $x_p = x_r + \frac{x_r - x_{r-1}}{v_r}(p - \gamma_r)$ (22.2).

За частния случай на медиана:

Определяме медианния интервал $[x_{r-1}, x_r]$, където γ_r е най-голямата честота, за която $\gamma_r \geq 0,5$, откъдето

$$Q_2 = x_r + \frac{d_r}{v_r}(0,5 - \gamma_r), \quad d_r = x_r - x_{r-1}.$$

Забележка 22.2. Освен за интервално, формули (22.2) могат да се използват и за неинтервално статистическо разпределение (ако $r=2$, то $v_r = v_2$ и трябва да се замени с γ_2).

Пример 22.5. Продължителността на разговорите по телефона в едно учреждение е представена чрез извадка с обем 100.

$(x_{i-1}, x_i]$	(0,2]	(2,4]	(4,6]	(6,8]	(8,10]
m_i	10	25	20	40	5

Да се намерят: а) квантилът $x_{0,2}$, б) долният и горният квантил Q_1 , Q_3 , в) медианата; г) 3%-процентилът $x_{3\%}$.

Решение. Използуваме таблицата

i	$(x_{i-1}, x_i]$	m_i	$v_i = \frac{m_i}{n}$	$\gamma_i = \frac{c_i}{n}$
1	(0,2]	10	0,1	0,1
2	(2,4]	25	0,25	0,35
3	(4,6]	20	0,20	0,55
4	(6,8]	40	0,40	0,95
5	(8,10]	5	0,05	1

а) Определяме интервала на квантила: $\gamma_2 = 0,35$ е най-малката относителна кумулативна честота, за която $0,2 < 0,35$. Следователно, $x_{0,2} \in (2,4]$ и се определя от израза (използваме втория ред)

$$x_{0,2} = x_2 + \frac{d_2}{v_2}(0,2 - \gamma_2) = 4 + \frac{2}{0,25}(0,2 - 0,35) = 4 - \frac{200}{25} \cdot 0,15 = 4 - 1,2 = 2,8$$

б) $0,25 < 0,35$. Следователно, $Q_1 = x_{0,25} \in (2,4]$ и

$$x_{0,25} = x_2 + \frac{d_2}{v_2}(0,25 - \gamma_2) = 4 + \frac{2}{0,25}(0,25 - 0,35) = 4 - \frac{200}{25} \cdot 0,1 = 4 - 0,8 = 3,2.$$

$0,75 < 0,95$. Следователно, $Q_3 = x_{0,75} \in (6,8]$ - (четвърти интервал) и

$$Q_3 = x_{0,75} = x_4 + \frac{d_4}{v_4}(0,75 - \gamma_4) = 8 + \frac{2}{0,40}(0,75 - 0,95) = 8 - \frac{20}{4} \cdot 0,2 = 8 - 1 = 7.$$

в) Q_2 е квантилът от ред $p = 0,5$, откъдето $0,5 < 0,55 \Rightarrow M_e \in (4,6]$,

$$Q_2 = x_3 + \frac{d_3}{v_3}(0,5 - \gamma_3) = 6 + \frac{2}{0,2}(0,5 - 0,55) = 6 - \frac{20}{2} \cdot 0,05 = 5,5$$

г) $x_{3\%} = x_{0,03}$. Тогава от $0,03 < 0,1$ следва, че $x_{0,03} \in (0,2]$ и

$$x_{0,03} = 2 + \frac{2}{0,1}(0,03 - 0,1) = 2 - 20 \cdot 0,07 = 2 - 1,4 = 0,6 \text{ .} \blacklozenge$$

Забележка. Всички квантили можеше да се намерят с помощта на

$$\text{кумулятивната функция } F^*(x) = \begin{cases} 0 & x \in (-\infty, 0] \\ 0,1 + 0,05(x-2) & x \in (0, 2] \\ 0,35 + 0,125(x-4) & x \in (2, 4] \\ 0,55 + 0,1(x-6) & x \in (4, 6] \\ 0,95 + 0,2(x-8) & x \in (6, 8] \\ 1 + 0,025(x-10) & x \in (8, 10] \\ 1 & x \in (10, \infty) \end{cases} \text{ и нейната}$$

графика. $\blacklozenge$

Пример 22.6. Ще намерим квантила $x_{0,4}$ на статистическото разпределение от пример 21.1.

Решение. Ако приложим формула (22.2), намираме първо, че $\gamma_3 = 0,5 > 0,4$, следователно (използваме третия ред на таблицата),

$$x_{0,4} = x_3 + \frac{x_3 - x_2}{\nu_3}(0,4 - \gamma_3) = 3 + \frac{1}{0,3}(0,4 - 0,5) = 3 - \frac{10}{3} \cdot 0,1 = 2\frac{2}{3}.$$

Ако приложим формула (22.1), то $np = 10 \cdot 0,4 = 4$, следователно, квантилът съвпада с петата варианта от вариационния ред, т.е. $x_{0,4} = 3$. $\blacklozenge$

Упражнения.

1. Времето, необходимо за почистване на дома (в минути) на 100 произволно избрани домакинства е следното:

(x_{i-1}, x_i)	(20,30)	(30,40)	(40,50)	(50,60)	(60,70)	(70,80)
m_i	10	12	25	42	7	4

а) Да се начертае диаграмата от тип бокс-плот.

б) Да намери до колко време отделят за почистване на дома 30% от домакинствата.

в) Повече от колко минути са необходими за почистване за 45% от домакинствата.

2. Секретарка е записала броя на набраните страници в работен ден в продължение на един месец (20 работни дни):

2, 5, 0, 3, 2, 6, 6, 10, 12, 2, 5, 13, 14, 0, 1, 2, 6, 5, 5, 6.

Да се намери модата, горният квартил, медианата и квантилът от ред 0,4.

3. В продължение на 100 дни контролор по качеството е записвал броят на дефектните изделия, които е проверил:

x_i	2	3	4	5	6	7	8	9	10
m_i	5	12	15	24	17	13	10	3	1

Да се намерят трите квартили на извадката, процентилът от ред 64, и критичната точка от ред 0,2.

§23. Средна, дисперсия и средно квадратично отклонение, начални и централни моменти на извадката.

В статистическите проучвания се търсят не точните параметри, на наблюдавания признак X , а техните статистически оценки. Затова голямо значение имат характеристиките, с които се получават различни средни стойности на извадката.

Нека е дадена извадката $\frac{X}{m_i} \mid \begin{matrix} x_1 & x_2 & \dots & x_k \\ m_1 & m_2 & \dots & m_k \end{matrix}$. От тези данни могат да се съставят следните числени характеристики:

$$\text{Средно аритметично } \bar{x} = \frac{1}{n} \sum_{i=1}^k x_i m_i. \quad (23.1)$$

Това е най-често използваната характеристика на извадката, затова обикновено я наричаме средна на извадката. Заедно с модата M_o и медианата M_e средната на извадката е една от трите характеристики на “центъра” на извадката.

Други важни характеристики са:

Дисперсия на извадката (извадъчна дисперсия)

$$s_x^2 = \frac{\sum (x_i - \bar{x})^2 m_i}{n} = \overline{x^2} - (\bar{x})^2, \quad (23.2)$$

$$\text{Извадъчно средно квадратично } s_x = \sqrt{s_x^2}, \quad (23.3)$$

които отразяват “разсейването” на наблюденията около средната на извадката.

Ако използваме относителните честоти ν_i , то

$$\bar{x} = \sum x_i \nu_i, \quad s_x^2 = \sum (x_i - \bar{x})^2 \nu_i$$

Вижда се приликата с формули (8.1) и (8.2) за математическо очакване и дисперсия на дискретна случайна величина. Затова свойствата на тези характеристики са подобни на свойствата на математическото очакване и дисперсията (виж §8), а именно, ако $c = \text{const}$ и са дадени две независими извадки на променливите X и Y , то

$$\begin{aligned} \overline{c} &= c & s_x^2 &= 0. \\ \overline{cx} &= c\bar{x}, & s_{(cx)}^2 &= c^2 s_x^2. \\ \overline{(x \pm y)} &= \bar{x} \pm \bar{y} & s_{x \pm y}^2 &= s_x^2 + s_y^2. \end{aligned}$$

Забележка 23.1. Ако данните не са групирани в статистическо разпределение, то $\bar{x}$ е просто сумата от всички варианти, разделена на броят им:

$$\bar{x} = \frac{1}{n}(x^{(1)} + \dots + x^{(n)}), \quad (23.4)$$

където $x^{(i)}$ е наблюдаваната стойност в i -тото измерване. Във формула (23.1) повтарящите се събираеми от (23.4) са обединени в произведението $m_i x_i$, защото в m_i е честотата на вариантите x_i

Аналогично, за извадъчната дисперсия на негрупирани данни имаме

$$s_x^2 = \frac{1}{n} \sum_{i=1}^n (x^{(i)} - \bar{x})^2. \quad (23.5)$$

Разглеждат се също начални и централни моменти:

Начален момент от ред p : $\overline{x^p} = \sum_{i=1}^k x_i^p m_i, \quad s=1,2,\dots$

Централен момент от ред p : $s_x^{(p)} = \sum_{i=1}^k (x_i - \bar{x})^p m_i, \quad s=2,3,\dots$

Между началните и централните моменти съществуват зависимости:

$$\begin{aligned} s_x^{(2)} &= s_x^2 = \overline{x^2} - (\bar{x})^2 \\ s_x^{(3)} &= \overline{x^3} - 3\bar{x}\overline{x^2} + 2(\bar{x})^3 \\ s_x^{(4)} &= \overline{x^4} - 4\bar{x}\overline{x^3} + 6\bar{x}^2\overline{x^2} - 3(\bar{x})^4 \end{aligned} \quad (23.6)$$

Ако величината X се мери в някакви мерни единици, то началните и централните моменти от ред p се измерват с p -тите им степени. Например, ако X се мери в метри, то средната и средно квадратичното също се мерят в метри, а дисперсията – в квадратни метри.

Въвеждат се също безразмерните характеристики

$$a_x = \frac{s_x^{(3)}}{(s_x)^3} - \text{асиметрия и } e_x = \frac{s_x^{(4)}}{(s_x)^4} - 3 - \text{ексцес.}$$

Други средни характеристики са:

Средно геометрично $G = \sqrt[n]{x_1^{m_1} x_2^{m_2} \dots x_k^{m_k}}.$

Средно хармонично $\frac{1}{H} = \frac{1}{n} \sum_{i=1}^k \frac{m_i}{x_i}.$

За негрупирани данни (в този случай полагаме $m_i = 1$)

$$G = \sqrt[n]{x_1 x_2 \dots x_n}, \quad \frac{1}{H} = \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}.$$

Ако извадката е зададена с интервален статистически ред, то всеки интервал $(x_{i-1}, x_i]$ се заменя със средата си x_i^* и се прилагат същите формули.

За улесняване на пресмятането им в много случаи е удобно полагане от вида (линейна трансформация)

$$u_i = ax_i - b,$$

откъдето, като имаме предвид, че $x_i = \frac{1}{a}(u_i + b)$, получаваме

$$\bar{x} = \frac{1}{a}[\bar{u} + b], \quad s_x^{(p)} = \frac{1}{a^p} s_u^{(p)}. \quad (23.7)$$

Пример 23.1. Дадена е извадката

(x_i, x_{i+1})	(1,3)	(3,5)	(5,7)	(7,9)	(9,11)
m_i	2	4	10	6	3

Да се намерят началните и централните моменти до четвърти ред, асиметрията a_s и ексцесът e_x .

Решение.

1) Всеки интервал заменяме със средата му $x_i^* = \frac{x_{i+1} + x_i}{2}$:

x_i^*	2	4	6	8	10
m_i	2	4	10	6	3

2) Вариантата с най-голяма честота е $x_3^* = 6$, затова пресмятанията ще се опростят, ако положим $u_i = \frac{1}{2}(x_i^* - 6)$. Така получаваме таблицата

u_i	m_i	$u_i m_i$	$u_i^2 m_i$	$u_i^3 m_i$	$u_i^4 m_i$
-2	2	-4	8	-16	32
-1	4	-4	4	-4	4
0	10	0	0	0	0
1	6	6	6	6	6
2	3	6	12	24	48
$\Sigma =$	25	4	30	10	90

в последния ред на която са резултатите от сумирането по съответните стълбове.

3) Пресмятаме началните моменти: $\bar{u} = \frac{1}{n} \sum_{i=1}^5 u_i m_i = \frac{4}{25} = 0,16,$

$$\overline{u^2} = \frac{1}{n} \sum_{i=1}^3 u_i^2 m_i = \frac{30}{25} = 1,2, \quad \overline{u^3} = \frac{1}{n} \sum_{i=1}^5 u_i^3 m_i = \frac{10}{25} = 0,4, \quad \overline{u^4} = \frac{1}{n} \sum_{i=1}^5 u_i^4 m_i = \frac{90}{25} = 3,6.$$

4) По формули (23.7) пресмятаме централните моменти:

$$s_u^{(2)} = \overline{u^2} - (\bar{u})^2 = 1,2 - 0,16^2 = 1,1744$$

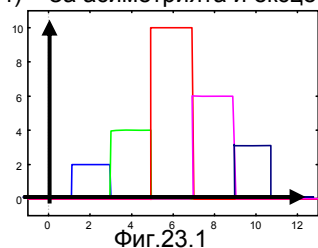
$$s_u^{(3)} = \overline{u^3} - 3\bar{u}\overline{u^2} + 2(\bar{u})^3 = 0,4 - 3 \cdot 0,16 \cdot 1,2 + 2 \cdot 0,16^3 = -0,1678$$

$$s_u^{(4)} = \overline{u^4} - 4\bar{u}\overline{u^3} + 6(\bar{u})^2\overline{u^2} - 3(\bar{u})^4 = 3,5263.$$

5) Пресмятаме всички моменти на изходната извадка като използваме формули (23.7). От равенството $u_i = \frac{1}{2}(x_i^* - 6)$ определяме, че $x_i^* = 2u_i + 6$, следователно, $\bar{x} = 2\bar{u} + 6 = 2.016 + 6 = 6,32$,

За централните моменти имаме $s_x^{(s)} = 2^s s_u^{(s)}$, откъдето $s_x^{(2)} = s_x^2 = 2^2 s_u^{(2)} = 4.1,1744 = 4,6976 \Rightarrow s_x = \sqrt{4,6976} = 2,1673$,
 $s_x^{(3)} = 2^3 s_u^{(3)} = 8.(-1,1678) = -1,3424$, $s_x^{(4)} = 2^4 s_u^{(4)} = 16.3,5263 = 56,4216$

4) За асиметрията и ексцеса получаваме



Фиг.23.1

$$a_x = \frac{-1,3424}{(s_x)^3} = -0,132,$$

$$e_x = \frac{s_x^{(4)}}{(s_x)^4} - 3 = -0,4427. \blacklozenge$$

Знакът на асиметрията е отрицателен, което означава, че данните са изместени на дясно (фиг. 23.1).

ОБРАБОТКА НА ДАННИ – ОСНОВНИ РЕЗУЛТАТИ

Означения: m_i , $v_i = \frac{m_i}{n}$, и $\gamma_i = \sum_{j=1}^i v_j$ - честота, относителна честота и

относителна кумулативна честота на вариантата x_i (на интервала (x_{i-1}, x_i) с номер i).

1) Неинтервално статистическо разпределение:

$$\begin{array}{c|cccc} X & x_1 & x_2 & \dots & x_k \\ \hline m_i & m_1 & m_2 & \dots & m_k \end{array}$$

Графично представяне – полигон.

$$\text{емпирична функция: } F^*(x) = \begin{cases} 0 & x \leq x_1 \\ \gamma_i & x_i < x \leq x_{i+1}, \quad i=1, \dots, k-1 \\ 1 & x > x_k \end{cases}$$

2) Интервално статистическо разпределение

$$\begin{array}{c|cccc} [x_{i-1}, x_i) & [x_0, x_1) & [x_1, x_2) & \dots & [x_{k-1}, x_k) \\ \hline m_i & m_1 & m_2 & \dots & m_k \end{array}$$

Графично представяне – хистограма.

$$\text{кумулятивна функция: } F(x) = \begin{cases} 0 & x \leq x_0 \\ \gamma_i + \frac{v_i}{x_i - x_{i-1}}(x - x_i) & x_{i-1} < x \leq x_i, \quad i=1, \dots, k-1 \\ 1 & x > x_k \end{cases}$$

3) Основни числени характеристики на извадка:

Квантил x_p от ред p : $x_p = x_r + \frac{x_r - x_{r-1}}{v_r}(p - \gamma_r)$, където е γ_r -

най-малката кумулативна честота, за която $p \leq \gamma_r$.

Средно аритметично $\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i m_i$.

извадъчна дисперсия $s_x^2 = \frac{\sum (x_i - \bar{x})^2 m_i}{n} = \overline{x^2} - (\bar{x})^2$, където

$$\overline{x^2} = \frac{1}{n} \sum_{i=1}^k x_i^2 m_i.$$

Общи задачи.

- Продължителността на 15 сърдечни трансплантации е както следва:
7,0 6,5 3,5 3,8 3,1 2,8 2,5 2,6 2,4 2,1 1,8 2,3 3,1 3,0 2,5.
а) Да се намерят медианата, долният и горният квантил и начертае бокс-плотът на извадката. Има ли случаи на изключения? б) Да се намери средната продължителност на една сърдечна трансплантация.
- Дадена е извадката
107 119 99 114 120 104 88 114 124 116 101 121 152 100 125 114 95 117.
а) Да се състави интервално разпределение с дължина на интервала $d=20$;
б) да определят модалният и медианният интервал на разпределението;
в) да се начертае хистограма на честотите;
г) да се намерят 30%-процентил, и квантилите от редове 0,2 и 0,6;
д) да се намерят извадъчните средна и дисперсия.
- При 20 посещения на магазин, студент е записал броя на чакащите пред него:
7, 7, 4, 5, 4, 6, 6, 9, 5, 5, 6, 6, 6, 7, 7, 6, 7, 6, 7, 9.
а) Да се начертае полигонът на честотите, да се намерят горният, долният квантил и медианата и начертае бокс-плотът на извадката.
б) да се намерят извадъчните асиметрия и ексцес.
- За изследване на неизвестната величина X е получена следната извадка
2, 3, 2, 0, 3, 2, 4, 1, 0, 2, 5, 5, 1, 4, 5, 1, 2, 1, 0, 0, 3, 6, 2, 1, 0
а) Да се намерят размахът и средата на извадката, медианата и горният квантил.
б) Да се начертае полигонът на относителните честоти.
в) Да се намерят извадъчните средна и дисперсия.
- Като се извърши подходяща линейна смяна, да се намерят моментите до трети ред на извадките:

$$\text{а) } \begin{array}{c|cccc} x_i & 15 & 16 & 17 & 18 & 19 \\ \hline m_i & 4 & 5 & 4 & 2 & 1 \end{array}$$

$$\text{б) } \begin{array}{c|cccccc} (x_{i-1}, x_i) & (10,14) & (14,18) & (18,22) & (22,26) & (26,30) & (30,34) \\ \hline m_i & 1 & 5 & 10 & 20 & 18 & 3 \end{array};$$

$$\text{в) } \begin{array}{c|cccccc} (x_{i-1}, x_i) & (34,36) & (36,38) & (38,40) & (40,42) & (42,44) & (44,46) \\ \hline m_i & 2 & 3 & 30 & 40 & 20 & 5 \end{array};$$

ОЦЕНКА НА ПАРАМЕТРИ

§24. Точкови оценки (статистики). Видове оценки – неизместена и изместена, състоятелна и ефективна.

Една от важните задачи на математическата статистика се състои по данните от извадката да се намерят някои от неизвестните параметри на изследваната случайна величина X .

Параметър на разпределението на случайна величина е неслучайно число, което, веднъж определено, остава неизменно за тази случайна величина.

Такива са всички числени характеристики на една случайна величина като математическо очакване, дисперсия, мода, медиана и т.н. Обикновено законът на една величина зависи от един (например, разпределението на Поасон), два (биномно $B(p, n)$, нормално $N(a, \sigma)$) или повече параметри (хипергеометрично).

Когато някой от параметрите на разпределението е неизвестен, то за определянето му се използват наличните данни от извадката като тук стоят три основни въпроса:

- по каква формула да се пресметне приближена стойност на параметъра;
- дали намерената стойност действително е оценка на параметъра;
- каква е точността на намерената оценка.

Нека е неизвестен параметърът θ на случайна величина X и нека за намирането му са направени n наблюдения като са наблюдавани стойностите $(x^{(1)}, \dots, x^{(n)})$, т.е. получена е извадка $(X_1, \dots, X_n)$ с обем n .

Точкова оценка θ^* на параметъра θ наричаме приближената стойност на този параметър, получена от извадката.

Оценката се нарича точкова, защото по определен начин от числата $(x^{(1)}, \dots, x^{(n)})$ сме получили само едно число (една точка от числовата права).

Очевидно, θ^* е функция на наблюдаваните стойности, т.е.

$$\theta^* = \theta^*(x^{(1)}, \dots, x^{(n)}).$$

Ще припомним, че:

- Резултатът $x^{(i)}$ от i -тото наблюдение, може да се тълкува като реализация на случайната величина $X_i = \{\text{наблюдавана стойност на } X \text{ в } i\text{-тото измерване}\} (i=1, \dots, n)$
- Предполагаме, че величините X_i са независими, приемат възможните стойности на величината X и имат закон, съвпадащ със закона на X .

Следователно, може да разглеждаме θ^* като реализация на случайната величина

$$\Theta^* = \Theta^*(X_1, \dots, X_n),$$

зависеща неслучайно (т.е. по точно определено правило) от величините $X_1, \dots, X_n$.

Произволна функция $f(X_1, \dots, X_n)$ се нарича статистика на извадката $(X_1, \dots, X_n)$. Изчислената стойност $f(x^{(1)}, \dots, x^{(n)})$ също се нарича статистика. Тя зависи от обема n на извадката и наблюдаваните стойности $x^{(1)}, \dots, x^{(n)}$ и се разглежда като реализация на $f(X_1, \dots, X_n)$.

Следователно, точковите оценки на генералната съвкупност (и всички получени в §§22-23 характеристики на извадката) са нейни статистики.

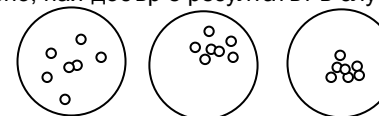
Най-простият начин за оценка на неизвестните параметри е неизвестната характеристика θ на генералната съвкупност да се оценява по съответстващата ѝ характеристика θ^* , взета от извадката (извадъчна характеристика). Например,

$\bar{x}$ е оценка за EX ,

s_x^2 е оценка за DX ,

$v_i = \frac{m_i}{n}$ е оценка за $p_i = P(X = x_i)$ и т.н.

Видове оценки. За да считаме, че изчислената стойност θ^* е добра оценка на неизвестния параметър θ , тя трябва да има определени качества. Например, ако резултатът от стрелба по мишена са представени на фиг. 24.1, то в случая а) казваме, че стрелбата е неефективна, в случая б) че е отместена (допуска се систематична грешка), докато в случая в) стрелбата е и ефективна, и неотместена. Естествено, най-добър е резултатът в случая в).



а) б) в)
Фиг.24.1.

Подобни качества за оценката θ^* на неизвестния параметър θ може да осигурим, ако случайната величина $\Theta^* = \Theta^*(X_1, \dots, X_n)$, отразяваща правилото, по което от извадката е изчислена оценката θ^* , притежава определени свойства.

Казваме, че оценката θ^* е неотместена, ако $E\Theta^* = \theta$.

Казваме, че оценката θ^* е отместена, ако $E\Theta^* \neq \theta$.

Пример 24.1. Тъй като $\bar{x} = \frac{1}{n}(x^{(1)} + \dots + x^{(2)})$ е наблюдаваната стойност на величината $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$ и $EX_1 = \dots = EX_n = EX$, то по свойства (15.1) имаме $E\bar{X} = \frac{1}{n}(EX_1 + \dots + EX_n) = EX$, т.е. средната на извадката $\bar{x}$ е неотместена оценка на математическото очакване EX (генералната средна).♦

Разликата $E\Theta^* - \theta$ се нарича отместване.

От свойствата на математическото очакване следва също, че:

- 1) Ако $E\Theta^* = \theta + a$, където $a = \text{const}$, то оценката $\Theta^* - a$ е неотместена.
- 2) Ако $E\Theta^* = a\theta$, то оценката $\frac{1}{a}\Theta^*$ е неотместена оценка.

Действително, $E(\Theta^* - a) = E\Theta^* - a = (\theta + a) - a = \theta$.

Аналогично се доказва свойство 2).

Както знаем, $E\Theta^*$ е средната стойност, около която се колебаят възможните стойности на Θ^* . Затова условието $E\Theta^* = \theta$ означава, че θ^* оценява неизвестната стойност на θ без систематическа грешка.

2. Казваме, че оценката θ^* е състоятелна, ако за произволно $\varepsilon > 0$
 $\lim_{n \rightarrow \infty} P(|\Theta^* - \theta| < \varepsilon) \rightarrow 1$.

Например, от теорема 16.3 за големите числа следва, че оценката $\bar{x}$ е състоятелна оценка на EX .

Ако една оценка е състоятелна, то колкото по-голям е обемът на извадката, толкова разликата $|\theta^* - \theta|$ е по-малка.

Пример 24.3. Ще покажем, че относителната честота $v_i = \frac{m_i}{n}$ е състоятелна оценка на вероятността $p_i = P(X = x_i)$.

Действително, v_i е относителната честота на събитието $X = x_i$ от n направени наблюдения. По теоремата на Бернули 16.4 за произволно $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{m_i}{n} - p_i\right| < \varepsilon\right) = 1,$$

следователно v_i е състоятелна оценка на вероятността p_i

3. За един и същ параметър θ на случайната величина X може да се получат няколко оценки от дадената извадка.

Казваме, че оценката θ^* е ефективна, ако за всяка друга оценка θ_1^* имаме $E(\Theta^* - \theta)^2 \leq E(\Theta_1^* - \theta)^2$

Ефективната оценка се характеризира с най-малкото разсейване около неизвестния параметър.

§25. Точкови оценки на генералните математическо очакване и дисперсия.

Да разгледаме една извадка $(X_1, X_2, \dots, X_n)$ на неизвестния признак X на генералната съвкупност, т.е. наблюдавани са случайните събития $X_i = x^{(i)}$, където случайните величини X_i имат разпределение, еднакво с разпределението на X .

Точкова оценка на математическото очакване EX .

Средната $\bar{x} = \frac{1}{n}(x^{(1)} + \dots + x^{(n)})$ на извадката е една от възможните

стойности на случайната величина $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$ и е оценка за математическото очакване на величината X .

За да определим свойствата на тази оценка, ще намерим:

$$E\bar{X} = \frac{1}{n}(EX_1 + \dots + EX_n) = \frac{1}{n}nEX = EX. \quad (25.1)$$

$$D\bar{X} = \frac{1}{n^2}(DX_1 + \dots + DX_n) = \frac{1}{n^2}nDX = \frac{1}{n}DX.$$

От полученото равенство $E\bar{X} = EX$ следва, че $\bar{x}$ е неотместена оценка на математическото очакване EX . От теорема 16.3 на Чебишев

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{X_1 + \dots + X_n}{n} - EX\right| < \varepsilon\right) = 1,$$

следва, че $\bar{x}$ е състоятелна оценка на EX .

Точкови оценки на дисперсията DX .

Оценка на дисперсията $DX = E(X^2) - (EX)^2$ на величината X е извадъчната дисперсия $s_X^2 = \frac{1}{n} \sum_{i=1}^n (x^{(i)} - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x^{(i)2} - (\bar{x})^2$, която е

наблюдаваната стойност на величината $S_X^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X})^2$ ($x^{(i)}$ е

резултатът от i -тото наблюдение). За да установим дали е отместена оценка, намираме математическото очакване на S_X^2 като използваме, че за еднакво разпределените величини X_i и X имаме $E(X_i^2) = E(X^2)$

$$E(S_X^2) = E\left[\frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X})^2\right] = \frac{1}{n} \sum_{i=1}^n E(X_i^2) - E(\bar{X}^2) = \frac{1}{n} \sum_{i=1}^n E(X^2) - E(\bar{X}^2).$$

Но $\frac{1}{n} \sum_{i=1}^n E(X^2) = \frac{1}{n} n E(X^2) = E(X^2)$, а като прибавим и извадим

$(EX)^2 = (E\bar{X})^2$ (виж формула (25.1)), получавме

$$E(S_X^2) = E(X^2) - (EX)^2 + (EX)^2 - E(\bar{X}^2) = DX - [E(\bar{X}^2) - (E\bar{X})^2] = DX - D\bar{X}.$$

Накрая, съгласно (25.2), имаме

$$E(S_X^2) = DX - \frac{1}{n} DX = \frac{n-1}{n} DX,$$

което означава, че извадъчната дисперсия s_x^2 е отместена оценка на дисперсията DX .

Не е трудно да се провери, че статистиката

$$\tilde{s}_x^2 = \frac{n}{n-1} s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

е неотместена оценка на дисперсията DX . Тя се нарича поправена извадъчна дисперсия.

Очевидно, $\tilde{s}_x^2 > s_x^2$, но за обем на извадката $n > 30$, разликата между двете числа е незначителна и може да приемем, че $\tilde{s}_x^2 \approx s_x^2$.

3. Оценка на средно квадратичното отклонение σ_X .

$$\text{Числото } \tilde{s}_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

се нарича поправено средно квадратично отклонение, стандарт или стандартна грешка и е неотместена оценка на средно квадратичното отклонение $\sigma = \sqrt{DX}$.

И тук за големи стойности на обема на извадката $\tilde{s}_x \approx s_x$.

Пример 25.1. Да се намерят точковите оценки на генералното математическо очакване EX и дисперсия DX на неизвестния параметър X по дадената извадка

x_i	0,01	0,04	0,08
m_i	5	3	2

Решение. Неотместена оценка за EX е средната на извадката

$$\bar{x} = \frac{1}{10} (0,01 \cdot 5 + 0,04 \cdot 3 + 0,08 \cdot 2) = 0,033.$$

За дисперсията DX имаме две оценки:

1) отместена оценка на DX е дисперсията на извадката $s_x^2 = \overline{x^2} - (\bar{x})^2$, където

$$\overline{x^2} = \frac{1}{10} (0,01^2 \cdot 5 + 0,04^2 \cdot 3 + 0,08^2 \cdot 2) = 0,00171 \Rightarrow s_x^2 = 0,00171 - 0,033^2 = 0,00071.$$

2) неотместена оценка на DX е

$$\tilde{s}_x^2 = \frac{n}{n-1} s_x^2 = \frac{10}{9} \cdot 0,00071 \approx 0,0008 \text{ (поправена дисперсия).}$$

Упражнения.

1. За изучаване на броя X на лицата в едно домакинство са получени данните:

x_i	1	2	3	4	5	6
m_i	8	17	11	10	2	2

Да се намерят точковите оценки на величината X .

2. Полученото месечно възнаграждение на 15 случайно избрани работници от дадено предприятие (в лв):

200, 200, 300, 300, 300, 400, 400, 400, 400, 400, 500, 500, 500, 500, 700.

Да се оцени по извадката средната работна заплата и отклонение от средната заплата в предприятието.

3. Дадена е следната извадка 0, 3, 2, 4, 1, 0, 2, 3, 2, 0, 1, 4, 5, 6, 2, 1, 2, 1, 0, 0, 3, 3, 2, 1, 0. Да се състави статистическото разпределение на извадката, да се намерят медианата и горният квантил и да се начертае полигонът на относителната честота и да се намерят точковите оценки на генералната съвкупност, от която е извлечена извадката.

4. Времето за подготовка на домашната работа за случайно избрани 80 ученика е както следва

(x_{i-1}, x_i)	(15, 20)	(20, 30)	(30, 40)	(40, 50)	(50, 60)
m_i	2	6	26	30	16

Да се намери средното време, необходимо за подготовка на домашна работа.

§26. Начини за намиране на оценки на параметри - метод на моментите. Метод на най-голямото правдоподобие.

Нека за изучаването на неизвестната случайна величина X е получена извадка с обем n . В §25 получихме точкови оценки за математическото очакване и дисперсията. Поставяме сега по-общия въпрос: как от получените данни да намерим оценки на други неизвестни параметри на разпределението на величината X .

Означаваме неизвестните параметри на разпределението с $\theta_1, \dots,$

θ_k (общ брой на неизвестните k). Ще разгледаме **метода на моментите** за намирането им, който използва факта, че неизвестните параметри могат да се изразят чрез моментите на разпределението. Например, за величината $X \sim B(p, n)$ (§11) е известно, че $EX = np$, $DX = npq = np(1-p)$. Следователно, ако търсим оценка на параметъра p ,

достатъчно е от извадката да намерим оценката $\bar{x}$ за математическото очакване EX . Тогава оценката за p е $p^* = \frac{\bar{x}}{n}$

Методът на моментите се състои в следното:

- 1) От извадката изчисляваме необходимите извадъчни моменти, които оценяват съответните моменти на разпределението на величината X .
- 2) Търсените оценки на параметрите $\theta_1, \dots, \theta_k$ се определят като се реши системата уравнения, отразяващи зависимостта между тези параметри и моментите на разпределението.

Пример 26.1. Известно е, че величината X е равномерно разпределена в интервала $[a, b]$ (§13). Да се намерят оценки за параметрите a и b по дадената извадка $\frac{(x_{i-1}, x_i)}{m_i} \mid \begin{matrix} (0,2) & (2,4) & (4,6) & (6,8) \\ 6 & 5 & 8 & 5 \end{matrix}$.

Решение: Ще използваме, че за равномерното разпределение имаме

$$E\xi = \frac{a+b}{2}, \quad D\xi = \frac{(b-a)^2}{12}.$$

Заменияйки интервалите със средите им, получаваме таблицата $\frac{x_i}{m_i} \mid \begin{matrix} 1 & 3 & 5 & 7 \\ 6 & 5 & 8 & 5 \end{matrix}$, от която за оценките $\bar{x}$ и $\tilde{s}_x^2$ на EX и DX получаваме:

$$\bar{x} = \frac{1}{24}(1.6 + 3.5 + 5.8 + 7.5) = 4;$$

$$s_x^2 = \overline{x^2} - (\bar{x})^2 = \frac{1}{24}(1.6 + 9.5 + 25.8 + 49.5) - 16 = \frac{14}{3} \Rightarrow \tilde{s}_x^2 = \frac{24}{23} s_x^2 = \frac{112}{23}$$

Следователно, за оценките a^* и b^* получаваме системата

$$\begin{cases} \frac{a^* + b^*}{2} = 4 \\ \frac{(a^* - b^*)^2}{12} = \frac{112}{23} \end{cases} \text{ с решение } a^* = 0,18, \quad b^* = 7,82. \blacklozenge$$

Един от най-разпространените методи за намиране на оценки на параметрите на разпределението на случайна величина е **методът на максималното правдоподобие**. Основната идея на този метод се състои в следното: ако $x^{(1)}, \dots, x^{(n)}$ са наблюдаваните стойности на величината X , да се намери такава функция, наречена **функция на правдоподобие** $L(\theta_1, \theta_2, \dots, \theta_k, x^{(1)}, \dots, x^{(n)})$ и зависеща от оценяваните параметри $\theta_1, \dots, \theta_k$, за която търсените оценки $\theta_1^*, \dots, \theta_k^*$ да са

стойностите, при които функцията приема максимална стойност, т.е. $L(\theta_1^*, \theta_2^*, \dots, \theta_k^*, x^{(1)}, \dots, x^{(n)}) = \max$.

Ще видим как се получава функцията на най-голямо правдоподобие за величина X , за определеност дискретна, с един неизвестен параметър θ на разпределението. Плътността на разпределение зависи от неизвестния параметър θ , т.е. $p_X = p_X(x, \theta)$, и тъй като наблюдаваните стойности $x^{(1)}, \dots, x^{(n)}$ са възможни стойности на величината, то за вероятностите на събитията $X = x^{(i)}$ имаме $P(X = x^{(i)}) = p_X(x^{(i)}, \theta)$, т.е. зависят от параметъра θ . Образоваме функцията

$$L(x^{(1)}, \dots, x^{(n)}, \theta) = P[(X = x^{(1)}) \cap \dots \cap (X = x^{(n)})] = \prod_{i=1}^N p_X(x^{(i)}, \theta),$$

за която променливата е θ , а $x^{(1)}, \dots, x^{(n)}$ са дадените числа.

Събитията $X = x^{(i)}$ са настъпили, затова най-достоверна ще е тази стойност θ^* на параметъра θ , за която вероятностите $P(X = x^{(i)})$ са възможно най-големи (най-близки до 1). Следователно като оценка на параметъра θ се приема тази стойност θ^* , за която

$$L(x^{(1)}, \dots, x^{(n)}, \theta) = \max, \text{ т.е. } \theta^* \text{ се определя от уравнението } \frac{dL}{d\theta} = 0.$$

Използва се също и логаритмичната функция на правдоподобие

$$\Lambda(x^{(1)}, \dots, x^{(n)}, \theta) = \ln L(x^{(1)}, \dots, x^{(n)}, \theta) = \sum_{i=1}^N \ln p_X(x^{(i)}, \theta),$$

за която θ^* е решение на по-удобното за получаване уравнение $\frac{d\Lambda}{d\theta} = 0$.

Пример 26.1. Да се намери по метода на най-голямото правдоподобие оценка на математическото очакване на величина, която има разпределение на Поасон.

Решение. Ако $X \sim Po(\lambda)$, то $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$, $k = 0, 1, 2, \dots$ и $EX = \lambda$.

Търсим оценка на неизвестния параметър $EX = \lambda$ по дадени наблюдавани стойности $x^{(1)}, \dots, x^{(n)}$. Образоваме логаритмичната функция на правдоподобие:

$$\Lambda(x^{(1)}, \dots, x^{(n)}, \lambda) = \sum_{i=1}^N \ln \frac{\lambda^{x^{(i)}} e^{-\lambda}}{x^{(i)}!} = \sum_{i=1}^N (x^{(i)} \ln \lambda - \lambda - \ln(x^{(i)}!))$$

Изчисляваме $\frac{d\Lambda}{d\lambda} = \sum_{i=1}^N \left(x^{(i)} \frac{1}{\lambda} - 1 \right) = \frac{1}{\lambda} \sum_{i=1}^N x^{(i)} - n$ и решаваме уравнението

$$\frac{1}{\lambda} \sum_{i=1}^N x^{(i)} - n = 0 \Rightarrow \lambda = \frac{1}{n} \sum_{i=1}^N x^{(i)}.$$

Следователно, получената по метода на най-голямото правдоподобие оценка на математическото очакване на величината $X \sim Po(\lambda)$ е отново извадъчната средна $\bar{x}$ ♦.

§27. Интервални оценки. Доверителен интервал и доверителна вероятност.

Нека θ е неизвестен параметър на случайната величина X и нека от извадка с обем n е изчислена нейна точкова оценка θ^* . Поставяме въпроса каква е точността на статистиката θ^* . Очевидно, колкото разликата $\theta^* - \theta$ е по-близка до нула или отношението $\frac{\theta^*}{\theta}$ е по-близко

до единица, толкова по-точна е оценката θ^* . Тези две нови статистики дават представа за точността на θ^* , но не са приложими, защото нямат точно определени закони на разпределение, а и самият параметър θ е неизвестен. С тяхна помощ, обаче, за всяка точкова оценка се намира подходяща статистика, законът на която е предварително известен.

Например, ако $X \sim N(EX, \sigma X)$, то за определяне на точността на оценката $\bar{x}$ на математическото очакване EX вместо статистиката $\bar{x} - EX$ се използва статистиката

$$Z = \frac{\bar{X} - EX}{\frac{\sigma}{\sqrt{n}}}, \quad \bar{X} = \frac{1}{n}(X_1 + \dots + X_n), \quad \sigma = \sigma X \quad (27.1)$$

разпределението на която, съгласно свойствата на нормално разпределените величини (виж теорема 13.1), е стандартното нормално разпределение.

Забележка 27.1. От централната гранична теорема и формули (15.3) следва, че при големи стойности на обема на извадката величината Z има разпределение, близко до стандартното нормално разпределение $N(0,1)$.

Следователно величината (27.1) има стандартно нормално разпределение ако:

- величината X има нормално разпределение, независимо от обема n на извадката;
- величината X има друго разпределение, но обемът n на извадката е голям.

По същия начин се въвеждат и други удобни статистики, които ще използваме по-нататък. По-важните от тях, както и теоретичните им разпределения са:

$$Z = \frac{\bar{x} - EX}{\frac{\sigma}{\sqrt{n}}} \rightarrow N(0,1) \quad (\text{Z-разпределение})$$

$$t = \frac{\bar{x} - EX}{\frac{\tilde{s}_x}{\sqrt{n}}} \rightarrow t(n-1) \quad (\text{t-разпределение на Стюдънт с } n \text{ степени на свобода})$$

$$\chi^2 = \frac{n\tilde{s}_x^2}{\sigma^2} \rightarrow \chi^2(n) \quad (\chi^2\text{-разпределение с } n-1 \text{ степени на свобода})$$

$$F = \frac{\tilde{s}_x^2/\sigma_x^2}{\tilde{s}_y^2/\sigma_y^2} = \frac{\tilde{s}_x^2/\tilde{s}_y^2}{\sigma_x^2/\sigma_y^2} \rightarrow F(n_1-1, n_2-1) \quad (\text{F-разпределение на Фишер с } n_1-1 \text{ и } n_2-1 \text{ степени на свобода}).$$

Определенията на цитираните теоретични разпределения са разгледани в §14.

С помощта на тези статистики се намира интервал, наречен доверителен, в който с дадена вероятност се намира истинската стойност на оценявания параметър.

Нека се оценява неизвестният параметър θ и нека е дадено число $0 < \gamma < 1$.

Доверителен интервал на параметъра θ с доверителна вероятност (надеждност) γ наричаме интервала (θ_1, θ_2) , за който $P(\theta_1 < \theta < \theta_2) = \gamma$.

Числото $\alpha = 1 - \gamma$ наричаме коэффициент на доверие, риск или ниво на значимост (на състоятелност). То е равно на вероятността изследваният параметър да се намира вън от доверителния интервал.

Радиусът $\delta = \frac{\theta_2 - \theta_1}{2}$ на доверителния интервал се нарича представителна грешка на оценката с доверителна вероятност γ .

Доверителната вероятност (надеждност) γ се избира в зависимост от конкретните условия и обикновено има стойности от порядъка на 0,9, 0,95, 0,99.

При дадена извадка с обем n намирането на доверителния интервал на параметъра θ по зададена доверителна вероятност γ се извършва по следния начин:

1. От извадката се изчислява точкова оценка θ^* на параметъра θ .
2. Избира се подходяща статистика $Y=Y(\theta^*,\theta)$, която да оценява разликата между θ^* и θ . Разпределението на тази величина трябва да е известно и да не зависи от параметъра θ .
3. В зависимост от зададената доверителна вероятност γ се определя интервала (y_1, y_2) , за който $P(y_1 < Y(\theta^*, \theta) < y_2) = \gamma$.
4. Границите на доверителния интервал на параметъра θ се определят като се реши относно θ неравенството $y_1 < Y(\theta^*, \theta) < y_2$.

§28. Доверителни интервали за математическото очакване и дисперсията на нормално разпределена случайна величина

Ще намерим доверителния интервал за математическото очакване $EX = \mu$ на **нормално разпределена величина** X , средно квадратичното отклонение на която е $\sigma X = \sigma$ и е известно.

Съгласно §27:

1) По дадената извадка с обем n намираме извадъчната средна $\bar{x}$, която е оценка на $EX = \mu$.

2) От забележка 27.1 следва, че $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$, т.е. разпределението

на случайната величина Z е известно и не зависи нито от $\bar{x}$, нито от μ . Следователно, Z може да се използва като статистика, оценяваща отклонението на величината от математическото ѝ очакване.

3) При доверителната вероятност γ можем да определим такова число ε , че вероятността на събитието $-\varepsilon < Z < \varepsilon$ да бъде равна на γ . Действително,

$$P(-\varepsilon < Z < \varepsilon) = P(|Z| < \varepsilon) = 2F(\varepsilon) - 1 = \gamma,$$

откъдето $F(\varepsilon) = \frac{1+\gamma}{2}$. Следователно, $\varepsilon = Z_{\frac{1+\gamma}{2}}$, т.е. ε е квантилът от ред

$\frac{1+\gamma}{2}$ на стандартното нормално разпределение.

4) Накрая, от неравенството $-\varepsilon < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < \varepsilon \Rightarrow -\varepsilon \frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < \varepsilon \frac{\sigma}{\sqrt{n}}$

получаваме, че $\bar{x} - Z_{\frac{1+\gamma}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{\frac{1+\gamma}{2}} \frac{\sigma}{\sqrt{n}}$ и това е доверителният интервал за математическото очакване.

Радиусът на интервала е $\delta = \varepsilon \frac{\sigma}{\sqrt{n}}$.

Аналогично се постъпва и в случая, когато не е известно средно квадратичното отклонение на X , но този път като статистика се използва величината $\frac{\bar{x} - EX}{\frac{\tilde{s}_x}{\sqrt{n}}}$, която има разпределение на Стьудънт

(§14) с $n-1$ степени на свобода. За определянето на доверителния интервал $(\bar{x} - \delta, \bar{x} + \delta)$, се пресмята квантилът от ред $\frac{1+\gamma}{2}$ на разпределението на Стьудънт с $n-1$ степени на свобода, които ще означаваме с $t_{\frac{1+\gamma}{2}}(n-1)$ и стойността му се взима от таблицата на стр. 169.

Така се получават следните правила за намиране на доверителни интервали:

Доверителен интервал за математическото очакване EX нормално разпределена величина X с доверителна вероятност γ по извадка с обем n :

$$\text{ако } \sigma X = \sigma \text{ е известно: } \bar{x} - Z_{\frac{1+\gamma}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{\frac{1+\gamma}{2}} \frac{\sigma}{\sqrt{n}}; \quad (28.1)$$

$$\text{ако } \sigma X \text{ не е известно: } \bar{x} - t_{\frac{1+\gamma}{2}}(n-1) \frac{\tilde{s}_x}{\sqrt{n}} < \mu < \bar{x} + t_{\frac{1+\gamma}{2}}(n-1) \frac{\tilde{s}_x}{\sqrt{n}}. \quad (28.2)$$

Понякога е необходимо да се определи какъв би трябвало да е обемът на извадката, че представителната грешка с дадена доверителна вероятност γ да бъде не по-голяма от дадено число δ_0 . Като се използва, че t -разпределението с $n=\infty$ съвпада с разпределението $N(0,1)$, се определя, че

$$n > \left(\frac{\sigma Z_{\frac{\gamma+1}{2}}}{\delta_0} \right)^2, \text{ ако } \sigma X = \sigma \text{ е известно, или } n > \left(\frac{\tilde{s}_x Z_{\frac{\gamma+1}{2}}}{\delta_0} \right)^2, \text{ ако } \sigma X \text{ не е известно и се оценява с } \tilde{s}_x.$$

Пример 28.1. Дадена е извадката

x_1	-2	1	2	3	4	5
m_1	2	1	2	2	2	1

.

а) Да се намерят доверителният интервал на математическото очакване EX и представителната грешка с доверителна вероятност $\gamma=0,95$. б) Колко трябва да е обемът на извадката, че представителната грешка за EX с доверителна вероятност 0,95 да е не по-голяма от 0,5?

Решение. От извадката изчисляваме точковите оценки на EX и DX :

$$\bar{x} = \frac{1}{n} \sum x_i m_i = 2, \quad \tilde{s}_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 m_i = \frac{52}{9}, \quad \tilde{s}_x \approx 2,4$$

а) Тъй като σX не е дадено, прилагаме формула (28.2). Изчисляваме $\frac{\gamma+1}{2} = \frac{1+0,95}{2} = 0,975$. От таблицата за t -разпределението при степени на свобода $n-1=10-1=9$ определяме квантила $t_{0,975}(9)=2,26$.

Следователно, представителната грешка е

$$\delta = t_{\frac{\gamma+1}{2}}(n-1) \cdot \frac{\tilde{s}_x}{\sqrt{n}} = 2,26 \cdot \frac{2,4}{\sqrt{10}} = 1,72,$$

а доверителният интервал за EX е $(0,8; 3,72)$.

б) За да определим обема на извадката, ще използваме отново формулата за δ , но този път ще вземем $t_{0,975}(\infty) = Z_{0,975} = 1,96$. Решаваме

неравенството $0,5 \geq 1,96 \cdot \frac{2,4}{\sqrt{n}}$, откъдето получаваме $n \geq 88$. ♦

По подобен начин се намират и доверителни интервали за дисперсията на нормално разпределена случайна величина, като в този случай се използва статистиката

$$\chi^2 = \frac{(n-1)\tilde{s}_x^2}{\sigma^2},$$

която има χ^2 -разпределение с $n-1$ степени на свобода (стр.170). Така се получават следните доверителни интервали за математическото очакване и дисперсията на нормално разпределена величина:

Доверителен интервал за дисперсията DX на нормално разпределена величина X с доверителна вероятност γ по извадка с обем n :

ако EX е известно: $\frac{ns_0^2}{\chi_{\frac{1+\gamma}{2}}^2(n)} < DX < \frac{ns_0^2}{\chi_{\frac{1-\gamma}{2}}^2(n)}, \quad s_0 = \frac{1}{n} \sum_{i=1}^k (x_i - \mu)^2 m_i$ (28.3)

ако EX не е известно: $\frac{(n-1)\tilde{s}_x^2}{\chi_{\frac{1+\gamma}{2}}^2(n-1)} < DX < \frac{(n-1)\tilde{s}_x^2}{\chi_{\frac{1-\gamma}{2}}^2(n-1)}$ (28.4)

Тук $\chi_{\frac{1+\gamma}{2}}^2(n)$ и $\chi_{\frac{1-\gamma}{2}}^2(n)$ са квантили на χ^2 -разпределението с n степени на свобода, а $\chi_{\frac{1+\gamma}{2}}^2(n-1)$ и $\chi_{\frac{1-\gamma}{2}}^2(n-1)$ - квантили на χ^2 -разпределението с $n-1$ степени на свобода

Пример 28.2. Да се намери доверителният интервал за DX с надеждност 0,95 по извадката от пример 28.1. ♦

Решение. Прилагаме формула (28.4). Изчисляваме квантилите на χ^2 -разпределението със степени на свобода $n-1=9$

$$\chi_{\frac{1+\gamma}{2}}^2(n-1) = \chi_{0,975}^2(9) = 19,05, \quad \chi_{\frac{1-\gamma}{2}}^2(n-1) = \chi_{0,025}^2(9) = 2,70$$

и определяме границите на доверителния интервал

$$\frac{9,2,4^2}{19,05} < DX < \frac{9,2,4^2}{2,70}.$$

Следователно, $2,73 < DX < 19,23$. ♦

Пример 28.3. По данни на администрацията на дадено предприятие средната продължителност на работната седмица е 40 часа. Избрани са 49 работника и за тях е изчислено, че работната седмица е средно 42 часа със средно отклонение 6 часа (поправено средно квадратично отклонение).

а) С каква доверителна вероятност може да приемем, че средната продължителност на работната седмица е 40 часа?

б) За по-нататъчни изследвания е необходимо да се определи средната седмична заетост с точност ± 1 час и доверителна вероятност 0,99. Колко работници трябва да съдържа извадката?

Решение.

а) Разликата между генералната средна $EX=40$ и средната $\bar{x}=42$ на извадката е равна на 2, т.е. 40 попада в доверителния интервал, ако радиусът му δ е по-голям от 2. От формулата за представителната грешка намираме

$$t_{\frac{1+\gamma}{2}}(9) = \frac{\delta \sqrt{n}}{\tilde{s}_x} > \frac{2 \cdot \sqrt{10}}{6} = 1,05.$$

От таблицата за квантилите определяме, че $t_{0,84}(9) \approx 1,05$.

Следователно, приетата средна продължителност попада в доверителния интервал, определен по дадената извадка, с доверителна вероятност, по-малка от 0,84.

б) Тук доверителната вероятност е $\gamma=0,99$. Ще изчислим какъв трябва да бъде обемът на извадката, че представителната грешка да бъде $\delta=1$.

Изчисляваме

$$Z_{\frac{1+\gamma}{2}} = Z_{0,995} = 2,575$$

и определяме n от неравенството $n > \left(\frac{Z_{\frac{1+\gamma}{2}} \cdot \tilde{s}_x}{\delta} \right)^2$, т.е.

$$n > \left(\frac{2,575.6}{1} \right)^2 \Rightarrow n > 238.7025.$$

Следователно, обемът на извадката трябва да е около 240. ♦

Упражнения.

1. За изследването на нормално разпределена случайна величина X е получена извадката

36 34 33 36 38 39 40 32.

Да се намерят точковите оценки на математическото очакване EX и дисперсията DX . Да се намерят доверителните интервали с доверителна вероятност $\gamma = 0,95$ за EX и DX .

2. Автомат пълни бутилки като съдържанието им по проект е 300 мл. Нека за определяне на средното количество μ в една бутилка за този автомат, са проверени 10 бутилки и са получени данните

299, 276, 283, 301, 297, 281, 300, 291, 295, 291.

а) Да се намери 90%-доверителен интервал за μ .

б) Ако съдържанието на бутилката е по-малко от 290, то тя се бракува. Какъв е процентът на бракуваните бутилки?

в) Ако съдържанието на бутилката е повече от 315, то предприятието търпи загуби. Какъв е процентът на препълнените бутилки?

3. Наблюдава се работата на автомат, който произвежда детайли по даден стандарт като за определяне на точността му са взети 30 детайла и са изчислени средната стойност $\bar{x} = 38$ на контролирания размер X и дисперсията на извадката

$$s_x^2 = 2,9.$$

а) коя от числените характеристики на извадката оценява точността на автомата?

б) да се намери доверителният интервал за точността на автомата с доверителна вероятност $\gamma = 0,95$.

4. При измерване на физична величина X са получени следните резултати:

-2, 0, -1, 1, 3.

а) да се намерят оценките за математическото очакване и дисперсията на величината X .

б) да се определи с надеждност $\gamma = 0,95$ доверителният интервал за математическото очакване на величината X .

5. В резултат на 5 измервания са получени резултатите

10, 9, 12, 15, 10.

Да се намери доверителен интервал на генералната дисперсия с доверителна вероятност $\gamma = 0,99$ за математическото очакване на измерваната величина.

6. От автомат за пакетиране на брашно са взети за контролно измерване 15 пакета и е изчислено $\bar{x} = 0,956g$, $\tilde{s}_x = 20g$. Да се оцени с надеждност $\gamma = 0,9$: а)

средното тегло на пакет брашно; б) точността на автомата;

7. Ако се приеме, че са дадени извадки от нормално разпределени величини, да се намерят доверителни интервали за:

а) средното време, необходимо за почистване на дома, по извадката от зад. 1, §22.

б) средната стойност и средното отклонение от тази стойност на броя X на дефектните изделия по данните от зад. 3, §22;

в) средната стойност и средното отклонение от тази стойност на броя X на пътниците в един автобус по данните от зад. 5, §21.

ПРОВЕРКА НА ХИПОТЕЗИ

§29. Статистически хипотези. Нулева и конкурираща хипотеза.

Грешка от първи и грешка от втори род. Статистически критерий

Ще се спрем на друга основна задача на математическата статистика – проверка на предположение (хипотеза) относно разпределението или параметър на разпределението на случайна величина X . Задачи от такъв характер са често срещани в различни области на науката, например, при обработка на научни резултати, прогнозиране на икономическото и общественото развитие и др. Решаването им преминава през следните етапи:

- 1) събиране на данни;
- 2) изказване на някаква хипотеза;
- 3) приемане и отхвърляне на хипотезата по утвърдени правила.

Нулева и конкурираща хипотеза.

Нека се изучава случайната величина X (дискретна или непрекъсната), разпределението на която е или неизвестно, или е неизвестен някой от параметрите на разпределението.

Всяко предположение за закона на разпределение или за параметрите на разпределението на случайната величина X се нарича статистическа хипотеза.

Статистическите хипотези се разделят на:

Прости - ако еднозначно определят разпределението на величината X . В противен случай хипотезата се нарича сложна.

Например, нека X има показателно разпределение. Хипотезата $\lambda = 3$ е проста, тъй като тогава плътността на разпределение е $p_X(x) = 3e^{-3x}$, $x > 0$. Хипотезата $0 < \lambda < 3$ е сложна хипотеза, тъй като законът на величината не е еднозначно определен.

Нека с H_0 е означена хипотезата, която ни интересува и която трябва да се провери (нулева хипотеза). Заедно с нея се разглежда и една от алтернативните ѝ хипотези H_1 , която се нарича конкурираща.

Конкуриращата хипотеза H_1 определя коя от хипотезите би трябвало да приемем, ако отхвърлим хипотезата H_0 .

Например, ако θ е неизвестен параметър на разпределението на величината X , то конкурираща хипотеза на хипотезата $H_0: \{\theta = \theta_0\}$ може да е една от хипотезите

$$H_1^{(1)}: \{\theta \neq \theta_0\}, H_1^{(2)}: \{\theta > \theta_0\} \text{ или } H_1^{(3)}: \{\theta < \theta_0\}.$$

Ще разглеждаме **само прости нулеви хипотези**.

Грешки от първи и втори род . Ниво на значимост.

Грешките, които може да направим са:

- да отхвърлим правилна хипотеза H_0 - грешка от първи род.
Вероятността да се допусне грешка от първи род се нарича ниво на значимост и се означава с α ;
- да приемем неправилна хипотеза H_0 - грешка от втори род.
Вероятността да се допусне грешка от втори род се означава с β .

Очевидно, трябва да се предполага, че α и β са малки числа.

Забележка 29.1. Дефинициите и смисъла на грешките от първи и втори род ще обясним с съпоставяне на решението за приемане на нулевата хипотеза с решението на съд за виновността на обвиняемия. Хипотезата H_0 е невинност на обвиняемия като обвинението представя доказателства за отхвърлянето ѝ. Ако доказателствата са убедителни (значими), то хипотезата за невинност се отхвърля. Един от най-важните принципи, от които се ръководи съдът при взимане на решение, е да не бъде осъден невинен човек, (отхвърляне на вярна хипотеза H_0), т.е. да не се допусне грешка от първи род.

На този принцип са организирани и изложените по-долу тестове за проверка на хипотези – заедно с конкуриращата хипотеза се определя и каква е допустимата вероятност α за отхвърляне на вярна хипотеза H_0 . Стойността на нивото на значимост α се избира от порядъка на 0,1, 0,05, 0,01 и т.н. По този начин се гарантира, че вероятността да се направи грешка е минимална.

Тест за приемане или отхвърляне на нулевата хипотеза. Нека са определени хипотезите H_0 и H_1 и нивото на значимост α като за проверка на хипотезата H_0 е взета извадка $(X_1, \dots, X_n)$ с обем n . Тъй като разполагаме единствено с числени данни, то тестовете за проверка са организирани така, че по определено правило от извадката да се получи едно число (наричано наблюдавана стойност) и в зависимост от стойността му да се вземе решение за приемане или отхвърляне на нулевата хипотеза. За целта:

1. Избира се подходяща статистика $\mathcal{K} = K(X_1, \dots, X_n)$, която да има предварително известно разпределение, т.е. разпределение, което не зависи от наблюдаваните стойности на признака X . Тогава изчислената по извадката стойност на $\mathcal{K}$ е една от нейните възможни стойности. Параметрите на разпределението на случайната величина $\mathcal{K}$ се наричат степени на свобода и се определят в предположение, че хипотезата H_0 е вярна.

Случайната величина $\mathcal{K}$, чрез която се проверява статистическа хипотеза се нарича статистически критерий или просто критерий. Формулата $\mathcal{K} = K(X_1, \dots, X_n)$, по която критерият се изразява чрез наблюдаваните стойности, се нарича статистика на критерия $\mathcal{K}$. Изчислената стойност на статистиката се означава с $K_{набл.}$ и се нарича наблюдавана стойност на критерия.

Пример 29.1 Нека X е нормално разпределена величина и от извадка с обем n сме изчислили извадъчната средна $\bar{x}$, която се явява наблюдавана стойност на величината $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$, която също е нормално разпределена, при това $E\bar{X} = EX$, $\sigma\bar{X} = \frac{\sigma X}{\sqrt{n}}$ (виж пример 15.1).

Тогава величината $Z = \frac{\bar{X} - EX}{\sigma X / \sqrt{n}}$ има стандартно нормално разпределение. Следователно,

- величината $Z \sim N(0,1)$, може да бъде статистически критерий,
- формулата $Z = \frac{\bar{X} - EX}{\sigma X / \sqrt{n}}$ е статистиката на критерия,
- наблюдаваната стойност е $Z_{набл.} = \frac{\bar{x} - EX}{\sigma X / \sqrt{n}}$ ♦

2. В зависимост от нивото на значимост α и конкуриращата хипотеза H_1 се определя

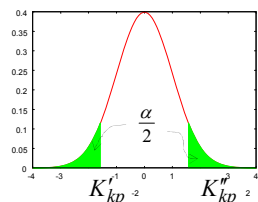
област D , за която ако $K_{набл.} \in D$ хипотезата се H_0 отхвърля. Тази област се нарича критична област.

Определянето на критичната област се основава на принципа, че малко вероятните събития са практически невъзможни. Тогава, ако хипотезата H_0 е вярна, то наблюдаваната стойност $K_{набл.}$ практически не попада в областите на малка вероятност на критерия $\mathcal{K}$. Следователно, имаме основание да приемем за критична такава област D , в която величината $\mathcal{K}$ приема стойности с малка вероятност, и тази вероятност да бъде равна на α . В случай, че $K_{набл.}$ попадне в областта D при вярна хипотеза H_0 , то ще направим грешка от първи род. Очевидно, $K_{набл.}$ е една от възможните стойности на $\mathcal{K}$ и $P(K_{набл.} \in D) = \alpha$.

Областта D , определена от стойностите на $K_{набл.}$, за които имаме основание да отхвърлим нулевата хипотеза, се нарича критична област.

Ако отхвърлим хипотезата H_0 трябва да приемем хипотезата H_1 . Затова освен от нивото на значимост, критичната област зависи и от конкуриращата хипотеза. Критичната област може да бъде:

Двустранна – ако $D = (-\infty, K'_{кр}) \cup (K''_{кр}, \infty)$:



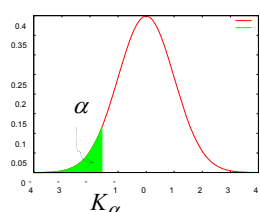
Фиг. 29.1.

където $K'_{кр} = K_{\frac{\alpha}{2}}$ и $K''_{кр} = K_{1-\frac{\alpha}{2}}$ са квантили от ред $\frac{\alpha}{2}$ и $1-\frac{\alpha}{2}$. Лицата

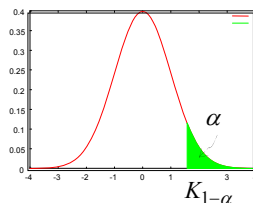
на двете застриховани области са равни на $\frac{\alpha}{2}$, а сумата им е равна на нивото на значимост α

Едностранна –

- ако $D = (-\infty, K_{кр})$ - лявостранна (фиг 29.2а)
- ако $D = (K_{кр}, \infty)$ - дясностранна (фиг 29.2б).



Фиг 29.2а.



Фиг 29.2б.

Тук критичните стойности са съответно квантилите K_{α} и $K_{1-\alpha}$ от ред α и $1-\alpha$, а лицата на застрихованите области са равни на α .

2. Взяма се решение за приемане или отхвърляне на хипотезата.

- Ако $K_{набл.} \notin D$, то нулевата хипотеза H_0 се приема.
- Ако $K_{набл.} \in D$, то нулевата хипотеза H_0 се отхвърля.

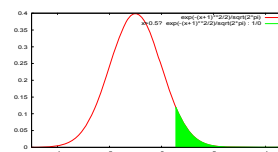
Връзка между вероятностите за грешка от първи и втори род. Мощност на критерия.

Нека са определени хипотезите H_0 и H_1 . Има две възможности да вземем правилно решение – когато приемем вярна хипотеза H_0 и когато отхвърлим грешна хипотеза H_0 , а всички случаи са дадени в таблицата

Решение \ Реалност	H_0 се приема	H_0 се отхвърля
H_0 е вярна	Правилно решение Вероятност: $1-\alpha$	Грешка от I-ви род Вероятност: α (ниво на значимост)
H_0 е грешна	Грешка от II-ри род Вероятност: β	Правилно решение Вероятност: $1-\beta$ (мощност на критерия)

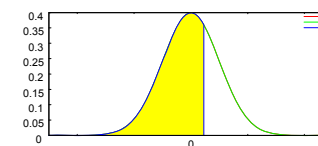
Вероятността да отхвърлим грешна хипотеза се нарича мощност на критерия.

Вероятностите α и β за грешки от първи и втори род са зависими една от друга. За да се види тази зависимост, да разгледаме графиките на плътностите на разпределението на критерия K съответно при условие, че H_0 е вярна (фиг. 29.3а) и при условие, че H_1 е вярна (фиг. 29.3б). Застрихованите области определят областта на грешно решение.



Графика на критерия, при условие, че хипотезата H_0 е вярна

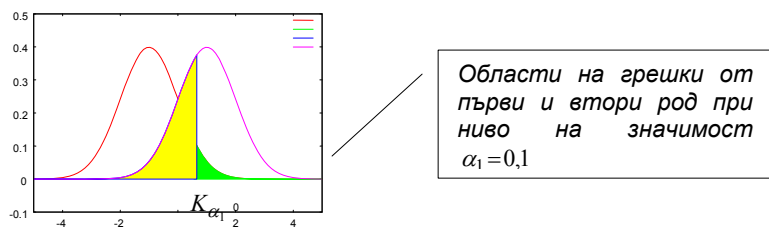
Фиг. 29.3а



Графика на критерия при условие че хипотезата H_1 е вярна

фиг.29.3б

На фиг. 29.4. са начертани областите на грешно решение при нива на значимост $\alpha=0.05$ и $\alpha=0.01$. Забелязваме, че с намаляване на вероятността от грешка от първи род, се увеличава вероятността от грешка от втори род.

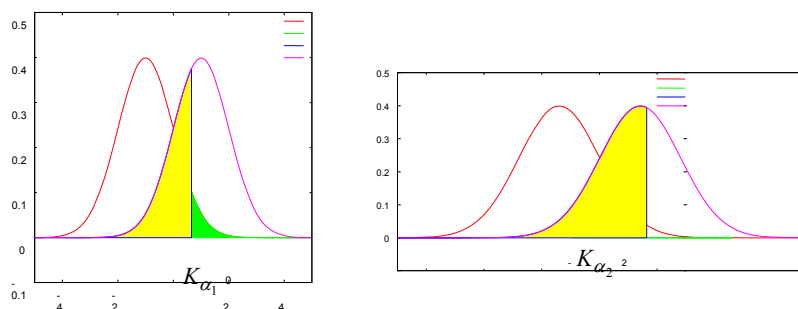


Фиг.29.4.

На фиг. 29.5. са съпоставени графиките на плътностите на критерия и областите на грешки от първи и втори род при различни стойности на обема на извадката.

- лявата при по-голям обем на извадката, а
- дясната — при по-малък.

Както се вижда, вероятността за грешки от първи и втори род намалява, при нарастване на обема на извадката.



Фиг.29.5.

Следователно,

- При намаляване на вероятността α на грешка от първи род се увеличава вероятността β за грешка от втори род и обратно.
- При увеличаване на обема на извадката вероятността за грешки от първи и втори род намалява.

Накрая, ще обобщим резултатите от тази глава:

- За всяка хипотеза H_0 е определен критерий K , и подходяща статистика, чрез която от извадката се пресмята наблюдаваната стойност $K_{набл.}$ на критерия.
- Параметрите на критерия K се наричат степени на свобода и в повечето случаи зависят от обема на извадката.
- Параметрите на статистиката на критерия се определят в предположение, че хипотезата H_0 е вярна.
- Критичната област (областта на отхвърляне на H_0) зависи от конкуриращата хипотеза H_1 и от нивото на значимост α .
- **Проверката на хипотеза H_0 при конкурираща хипотеза H_1 и ниво на значимост α се извършва по следния начин:**
 - 1) От извадката се пресмята $K_{набл.}$.
 - 2) В съответствие с H_1 , нивото на значимост α и броя на степените на свобода се определя критичната област D .
 - 3) Взяма се решение: ако $K_{набл.} \notin D$ - приема се H_0
ако $K_{набл.} \in D$ - отхвърля се H_0

§30. Непараметрични тестове - χ^2 - критерий на Пирсън за вида на разпределението.

Хипотезите и тестове за приемането им може да разделим на две групи :

Параметрични - хипотези, които се отнасят за някои количествени параметри на разпределението на изучавания признак X ,
Непараметрични - хипотези относно други характеристики на признака X .

Към непараметричните хипотези спадат хипотезите относно вида на разпределението, например, хипотезата $H_0 = X \sim B(N, p)$. Проверката за верността на хипотезата H_0 се извършва на базата на извадка с обем n :

$$\frac{X}{m_i} \mid \begin{matrix} x_1 & x_2 & \dots & x_k \\ m_1 & m_2 & \dots & m_k \end{matrix}, \quad n = \sum_{i=1}^k m_i.$$

Преди проверката, от извадката понякога се намират оценки на необходимите параметри на разпределението (за биномното разпределение това са параметрите N и p).

Нека броят на определените от извадката параметри на разпределението е l .

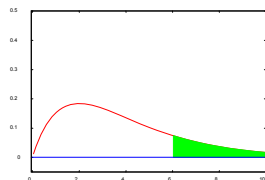
След като е определен какъв е видът на разпределението (например, биомно разпределение) и какви са параметрите на това

разпределение (параметрите N и p), се пристъпва към сравняване на резултатите от наблюденията с теоретичните резултати, изчислени съгласно издигнатата хипотеза. Разглеждат се разликите $m_i - m'_i$ между наблюдаваните честоти m_i и теоретичните честоти m'_i – честотите, които би трябвало да имат вариантите, ако хипотезата H_0 е вярна. Използва се статистиката

$$\chi^2_{\text{набл}} = \sum_{i=1}^k \frac{(m_i - m'_i)^2}{m'_i},$$

която има χ^2 -разпределение със степени на свобода $s = k - l - 1$.

Критичната област е дясностранна (защрихованата област на фиг. 30.1), определена от критичната стойност от $\chi^2_{\alpha}(s)$ ред α на случайната величина $\chi^2(s)$ (равна на квантила $\chi^2_{1-\alpha}(s)$ от ред $1-\alpha$).



Фиг. 30.1

Границите за приемане или отхвърляне се определят от избраното ниво на значимост. По този начин:

Проверката на хипотезата H_0 за вида на разпределението при ниво на значимост α се превежда по следния начин:

1) намират се теоретичните честоти m'_i и наблюдаваната стойност

$$\chi^2_{\text{набл}} = \sum_{i=1}^k \frac{(m_i - m'_i)^2}{m'_i}$$

2) изчисляват се степените на свобода $s = k - l - 1$, където:

- k е броят на различните варианти в извадката,
- l е броят на параметрите на разпределението, оценките на които са определени от извадката

3) от таблицата се взима $\chi^2_{\text{кр.}} = \chi^2_{1-\alpha}(s)$, където $\chi^2_{1-\alpha}(s)$ е квантилът от ред $1-\alpha$ на χ^2 -разпределението със степени на свобода s .

4) Взима се решение:

Ако $\chi^2_{\text{набл.}} < \chi^2_{\text{кр.}}$, то хипотезата H_0 се приема

Ако $\chi^2_{\text{набл.}} \geq \chi^2_{\text{кр.}}$, то H_0 се отхвърля

Изчисляване на теоретичните честоти. Нека $p_X(x)$ и $F_X(x)$ са съответно плътността и функцията на разпределение на величината X съгласно издигнатата хипотеза H_0 . Теоретичните честоти изчисляваме по формулата

$$m'_i = np'_i \quad (i=1, \dots, k),$$

където са теоретични честоти, които пресмятаме в зависимост от разпределението на извадката по следния начин:

- Ако статистическото разпределение е неинтервално, т.е.

извадката е дадена във вида $\frac{X}{m_i} \left| \begin{array}{c} x_1 \\ m_1 \end{array} \right| \frac{x_2}{m_2} \dots \frac{x_k}{m_k}$, то

$$p'_i = P(X = x_i) = p_X(x_i) \quad (i=1, \dots, k).$$

- Ако статистическото разпределение е интервално, т.е. извадката е дадена примерно във вида

$\frac{[x_1, x_{i+1})}{m_i} \left| \begin{array}{c} [x_1, x_2) \\ m_1 \end{array} \right| \frac{[x_2, x_3)}{m_2} \dots \frac{[x_k, x_{k+1})}{m_k}$, то

$$p'_i = P(x_i < X < x_{i+1}) = F_X(x_{i+1}) - F_X(x_i) \quad (i=1, \dots, k).$$

Пример 30.1. Проверява се хипотезата H_0 , че 41% от хората имат кръвна група А, 9% - кръвна група В, 46% - нулева кръвна група и 4% - кръвна група АВ. Да се провери тази хипотеза с ниво на значимост $\alpha=0,02$, ако са получени следните данни:

кръвна група	A	B	O	AB
m_i	74	25	26	15

Решение. 1) Тъй като извадката има обем 200, то съгласно хипотезата,

$$m'_1 = np'_1 = 0,41 \cdot 200 = 82, \quad m'_2 = 0,09 \cdot 200 = 18,$$

$$m'_3 = 0,46 \cdot 200 = 92, \quad m'_4 = 0,04 \cdot 200 = 8,$$

откъдето

$$\chi^2_{\text{набл}} = \sum_{i=1}^4 \frac{(m_i - m'_i)^2}{m'_i} = \frac{(74-82)^2}{82} + \frac{(25-18)^2}{18} + \frac{(26-92)^2}{92} + \frac{(15-8)^2}{8} = 10,02$$

2) От извадката не сме изчислявали параметри на разпределението, следователно, броят на степените на свобода е $s = 4 - 0 - 1 = 3$

3) Определяме $\chi^2_{\text{кр.}} = \chi^2_{1-\alpha}(s) = \chi^2_{1-0,02}(3) = \chi^2_{0,98}(3) = 7,81$.

4) Тъй като $\chi^2_{\text{набл.}} > \chi^2_{\text{кр.}}$, то хипотезата се отхвърля. ♦

Забележка 30.1. Ще обърнем внимание, че в този пример беше разгледана не количествена, а качествена величина, която приема 4 различни "стойности" А, В, О и АВ.

Пример 30.2. При ниво на значимост $\alpha=0,1$ да се провери хипотезата, че величината X , за която е получена извадката

$\frac{(x_{i-1}, x_i)}{m_i} \left| \begin{array}{c} (0,2) \\ 6 \end{array} \right| \frac{(2,4)}{5} \frac{(4,6)}{8} \frac{(6,8)}{5}$, има равномерно разпределение.

Решение.

1) **Определяне на параметрите на разпределението.** В пример 26.1 получихме оценки $a^*=0,18$ и $b^*=7,82$ за интервала $[a, b]$ от стойности на величината X .

2) **Пресмятане на $\chi^2_{набл}$.** Плътността на разпределението на равномерно разпределена величина в интервала $[0,18; 7,82]$ е

$$p_X(x) = \frac{1}{7,82-0,18} = 0,131 \text{ за } x \in [0,18; 7,82] \text{ и при } x \notin [0,18; 7,82], \text{ а}$$

функцията на разпределение е

$$F_X(x) = \begin{cases} 0 & x < 0,18 \\ 0,131x - 0,024 & 0,18 \leq x \leq 7,82 \\ 1 & x > 7,82 \end{cases}$$

Изчисляваме теоретичните вероятности:

$$p'_1 = P(0 < X \leq 2) = F_X(2) - F_X(0) = 0,238,$$

$$p'_2 = P(2 < X \leq 4) = F_X(4) - F_X(2) = 0,500 - 0,238 = 0,262,$$

$$p'_3 = P(4 < X \leq 6) = F_X(6) - F_X(4) = 0,262, \quad p'_4 = P(6 < X \leq 8) = F_X(8) - F_X(6) = 0,238.$$

Умножавайки по обема $n=24$, получаваме теоретичните честоти. Пресмятанятия подреждаме в таблица:

$(x_{i-1}, x_i]$	m_i	$m'_i = np'_i$	$\frac{(m_i - m'_i)^2}{m'_i}$
(0, 2]	6	5,71	0,014
(2, 4]	5	6,28	0,261
(4, 6]	8	6,28	0,828
(6, 8]	5	5,71	0,089
$\Sigma =$			1,192

. Следователно, $\chi^2_{набл.} = 1,192$.

3. **Изчисляване на $\chi^2_{кр}$.** При определянето на степените на свобода имаме предвид, че интервалите на извадката са 4 и че от тази извадка са определени два параметра на разпределението – границите интервала $[0,18; 7,82]$, т.е. $s = k - l - 1 = 4 - 2 - 1 = 1$, Следователно, при $\alpha = 0,1$

$$\chi^2_{кр} = \chi^2_{1-0,1}(1) = \chi^2_{0,9}(1) = 2,706.$$

4. **Решение за приемане или отхвърляне на хипотезата.** Тъй като $\chi^2_{набл.} < \chi^2_{кр.}$, то няма основание за отхвърляне на хипотезата за равномерно разпределение на величината X .

Пример 30.3. Контролира се размерът X на детайлите, изработвани на струг. С ниво на значимост $\alpha = 0,1$ да се провери хипотезата за нормално разпределение на контролирания размер.

Решение. 1) С помощта на таблицата от извадката изчисляваме оценката $\bar{x}$ на $a = EX$ и оценката $\tilde{s}_x$ на $\sigma = \sigma X$:

(x_{i-1}, x_i)	m_i	$x^* = (x_i + x_{i-1})/2$	$m_i x_i$	$m_i x_i^2$
(3, 4)	5	3,5	17,5	61,25
(4, 5)	15	4,5	67,5	303,75
(5, 6)	23	5,5	126,5	695,75
(6, 7)	19	6,5	123,5	802,75
(8, 9)	6	7,5	45	337,5
$\Sigma =$	68	27,5	380	2201

Получаваме: $\bar{x} = 5,5882$, $\tilde{x}^2 = 32,3676$, $\tilde{s}_x^2 = 1,1563$, $\tilde{s}_x = 1,0753$.

2) За изчисляване на теоретичните честоти и $\chi^2_{набл}$ използваме таблицата

(x_{i-1}, x_i)	m_i	$p_i = F\left(\frac{x_i - \bar{x}}{\tilde{s}_x}\right)$	$p_{i-1} = F\left(\frac{x_{i-1} - \bar{x}}{\tilde{s}_x}\right)$	$m'_i = n(p_i - p_{i-1})$	$\frac{(m_i - m'_i)^2}{m'_i}$
(3, 4)	5	0,069836	0,008042	4,201982	0,151555
(4, 5)	15	0,292175	0,069836	15,11904	0,000937
(5, 6)	23	0,649114	0,292175	24,27182	0,066642
(6, 7)	19	0,905391	0,649114	17,42689	0,142004
(8, 9)	6	0,987547	0,905391	5,586606	0,03059
$\Sigma =$					0,391729

Следователно, $\chi^2_{набл.} = 0,3917$.

3) Изчисляваме $\chi^2_{кр} = \chi^2_{1-0,1}(5-2-1) = \chi^2_{0,9}(2) = 4,61$

4) Тъй като $\chi^2_{набл.} < \chi^2_{кр.}$, то приемаме хипотезата за нормално разпределение.

Упражнение.

1. По извадката $\frac{x_i}{m_i} \begin{matrix} 1 & 2 & 3 & 4 & 5 \\ 110 & 130 & 70 & 90 & 100 \end{matrix}$, може ли да се твърди с вероятност 0,01, че дискретният признак X , които приема възможни стойности 1,2,3,4 или 5, е равномерно разпределен?

§31. Параметрични тестове - тестове за математическото очакване и дисперсията на нормално разпределена генерална съвкупност.

Параметричните тестове проверяват хипотези относно стойността на параметър θ на разпределението на изследван количествен признак X на генералната съвкупност. Както вече отбелязахме, предполагаме, че хипотезата H_0 е проста, т.е. $H_0 = \{\theta = \theta_0\}$. Като конкурираща може да се приеме една от хипотезите $H_1^{(1)}: \{\theta \neq \theta_0\}$, $H_1^{(2)}: \{\theta > \theta_0\}$ или $H_1^{(3)}: \{\theta < \theta_0\}$,

от която зависи вида на критичната област.

Пример 31.1. Да разгледаме нормално разпределена случайна величина X с неизвестно математическо очакване и известно средно квадратично отклонение $\sigma=1$. Нека от извадката $(X_1, \dots, X_9)$ с обем $n=9$ е изчислена извадъчната средна $\bar{x}=2,5$. Ще проверим хипотезата $H_0: \{EX=3\}$ при ниво на значимост $\alpha=0,1$ и при конкурираща хипотеза:

а) $H_1 = \{EX \neq 3\}$; б) $H_1 = \{EX > 3\}$; в) $H_1 = \{EX < 3\}$.

Решение. Ако допуснем, че хипотезата H_0 е вярна, то величината X и величините $X_1, \dots, X_n$ (резултатите от наблюденията на X) са разпределени по закона $N(3,1)$, т.е. те са нормално разпределени величини с математическо очакване $EX=EX_i=3$ и $\sigma X=\sigma X_i=1$ ($i=1, \dots, 9$). Тогава $\bar{x}=2,5$ е стойност на величината $\bar{X}=\frac{1}{9}(X_1+\dots+X_9)$, която също е нормално разпределена.

Съгласно свойствата на математическото очакване и дисперсията

$$E\bar{X}=\frac{1}{9}(EX+\dots+EX)=\frac{1}{9}9.3=3, \quad D\bar{X}=\frac{1}{9^2}(DX+\dots+DX)=\frac{1}{81}9.1=9.1=\frac{1}{9},$$

$$\sigma\bar{X}=\sqrt{D\bar{X}}=\sqrt{\frac{1}{9}}=\frac{1}{3}.$$

Следователно, ако H_0 е вярна, то $\bar{X} \sim N\left(3, \frac{1}{9}\right)$ и $\bar{x}=2,5$ е една от нейните възможни стойности. По тази стойност ние трябва да приемем или отхвърлим H_0 при ниво на значимост α и при конкурираща хипотеза H_1 .

Тъй като $\bar{X}$ е неотместена оценка на EX , то е малко вероятно наблюдаваната стойност $\bar{x}$ на $\bar{X}$ да се различава съществено от EX . Следователно, може да приемем, че хипотезата H_0 не е вярна само, ако $\bar{x}$ попадне в област с малка вероятност, т.е. в краищата на реалната права $(-\infty, \infty)$. По този начин ще бъде малко вероятно да отхвърлим вярна хипотеза H_0 и да приемем невярната хипотеза H_1 . Следователно, критичната област D ще определим в зависимост от нивото на значимост α и от конкуриращата хипотеза H_1 по следния начин:

а) $H_1 = \{EX \neq 3\}$, т.е. $H_1 = \{3 < EX \text{ или } 3 > EX\}$ — критичната област съдържа възможни стойности както по-големи, така и по-малки от 3, т.е. стойности на $\bar{x}$, за които $|\bar{x}-3|$ е голямо число.

Критичната област ще бъде определена, ако намерим такова число δ , че вероятността за събитието $\{|\bar{x}-3| > \delta\}$ да бъде равна на нивото на значимост $\alpha=0,1$, т.е. $P(\text{да приемем хипотезата } H_1) = P(|\bar{x}-EX| > \delta) = 0,1$. Но $P(|\bar{x}-EX| > \delta) = 1 - P(|\bar{x}-EX| \leq \delta) \Rightarrow P(|\bar{x}-EX| \leq \delta) = 1 - 0,1 = 0,9$.

Като използваме формула (14.2) за величината $\bar{X} \sim N\left(3, \frac{1}{9}\right)$, за δ

$$\text{получаваме } P(|X-EX| \leq \delta) = 2F\left(\frac{\delta}{\sigma}\right) - 1 = 0,9 \Rightarrow 2F(3\delta) = 1,9 \Rightarrow F(3\delta) = 0,95.$$

От таблицата за функцията $F(x)$ определяме, че $3\delta = 1,645 \Rightarrow \delta = 0,548$.

Тогава, областта за приемане на нулевата хипотеза е интервалът $[2,452; 3,548]$, а областта на отхвърляне е двустранната област $D = (-\infty, 2,452) \cup (3,548, \infty)$. На черт. 29.1 са представени двете области като $K'_{кр.} = 2,452$ и $K''_{кр.} = 3,548$. Лицето на заштрихованите области е равно на нивото на значимост $\alpha=0,1$. Тъй като $\alpha \notin D$, то нямаме основание да отхвърлим хипотезата H_0 .

б) $H_1 = \{EX > 3\}$. В този случай областта на отхвърляне на H_0 се определя от равенството $P(\bar{x} - 3 > \delta) = 0,1$, т.е. $P(\bar{x} > \delta + 3) = 0,1$. Но

$$P(\bar{x} > \delta + 3) = 1 - P(\bar{x} < \delta + 3) = 1 - F\left(\frac{\delta + 3 - 3}{1/3}\right) = 1 - F(3\delta)$$

$$\text{Следователно, } F(3\delta) = 0,9 \Rightarrow 3\delta = 1,285 \Rightarrow \delta = 0,43.$$

От тук определяме критичната област $D = (\delta + 3, \infty) = (3,43, \infty)$ и областта на приемане $(-\infty, 3,48]$ (фиг. 29.2б).

Тъй като $x = 2,5 \in (-\infty, 3,48)$, то и в този случай приемаме хипотезата H_0 .

в) $H_1 = \{EX < 3\}$. Областта на отхвърляне на H_0 определяме от равенството

$$P(\bar{x} - 3 < \delta) = 0,1 \Rightarrow P(\bar{x} < \delta + 3) = F\left(\frac{\delta + 3 - 3}{1/3}\right) = F(3\delta) = 0,1.$$

Тъй като $0,1 < 0,5$, то $3\delta < 0$ и за да намерим δ , ще използваме, че

$$F(3\delta) = 1 - F(-3\delta), \text{ т.е. } F(-3\delta) = 1 - 0,1 = 0,9.$$

От таблицата получаваме $-3\delta = 1,285$. Следователно, $\delta = -0,43$, а критичната област е $D = (-\infty, \delta + 3) = (-\infty, 2,57)$ (фиг. 29.2а).

Тъй като $\bar{x} = 2,5 \in (-\infty, 2,57)$, то нямаме основание да приемем хипотезата $H_0: \{EX=3\}$ при конкурираща хипотеза $H_1 = \{EX < 3\}$. ♦

Както се вижда от примера, при едни и същи резултати от наблюденията една хипотеза може да бъде приета или отхвърлена в зависимост от конкуриращата хипотеза. ♦

В разгледания пример, за проверка на хипотезата $H_0: \{EX=3\}$ използвахме, че ако хипотезата е вярна, то $\bar{X} \sim N(3, 1/3)$. Очевидно, знаем също разпределението на статистиката $Z = \frac{\bar{X}-3}{1/3} \sim N(0,1)$. Намерените критични стойности са квантилите съответно от ред $1-\alpha/2=0,95$, $1-\alpha=0,9$ и $\alpha=0,1$ на стандартното нормално разпределение. Следователно, за хипотезата $H_0 = \{EX=a\}$, подходяща за статистически критерий е величината $Z \sim N(0,1)$.

Ако средно квадратичното отклонение не е известно, то от извадката пресмятаме неговата оценка – поправеното средно квадратично отклонение $\tilde{s}_x$. Тогава като статистика на критерия за проверка на хипотезата $H_0 = \{EX=a\}$ се използва величината $\frac{\bar{X}-a}{\tilde{s}_x/\sqrt{n}}$.

Доказва се, че тази величина има разпределение $t(n-1)$ на Стьудънт (t -разпределение с $n-1$ степени на свобода. Така обобщаваме:

Тестове за проверка на хипотезата $H_0 = \{EX=a\}$ относно математическото очакване EX на нормално разпределена величина с ниво на значимост α :

- **При известно σX : Статистика:** $Z_{набл.} = \frac{\bar{x}-a}{\sigma X/\sqrt{n}} \sim N(0,1)$. (31.1)

Критична област: ако $H_1 = \{EX \neq a\}$, то $|Z_{набл.} - a| > Z_{1-\frac{\alpha}{2}}$;

ако $H_1 = \{EX > a\}$, то $Z_{набл.} > Z_{1-\alpha}$;

ако $H_1 = \{EX < a\}$, то $Z_{набл.} < Z_{\alpha}$.

- **При неизвестно σX : Статистика:** $t_{набл.} = \frac{\bar{x}-a}{\tilde{s}_x/\sqrt{n}} \sim t(n-1)$. (31.2)

Критична област: ако $H_1 = \{EX \neq a\}$, то $|t_{набл.} - a| > t_{\frac{\alpha}{2}}$;

ако $H_1 = \{EX > a\}$, то $t_{набл.} > t_{1-\alpha}$;

ако $H_1 = \{EX < a\}$, то $t_{набл.} < t_{\alpha}$.

По подобен начин са организирани и тестовете за проверка на хипотезите, касаещи дисперсията. Тук като статистически критерий се използва χ^2 -разпределението с n или $n-1$ степени на свобода:

Тестове за проверка на хипотезата $H_0 = \{DX = c^2\}$ относно дисперсията DX на нормално разпределена величина с ниво на значимост α :

- **При известно $EX = \mu$: Статистика:**

$$\chi_{набл.}^2 = \frac{ns_0^2}{c^2} \sim \chi^2(n), \quad \text{където } s_0^2 = \frac{\sum (x_i - \mu)^2 m_i}{n}.$$

Критична област: ако $H_1 = \{DX \neq c^2\}$, то $\chi_{набл.}^2 < \chi_{\frac{\alpha}{2}}^2(n)$, $\chi_{набл.}^2 > \chi_{1-\frac{\alpha}{2}}^2(n)$;

ако $H_1 = \{DX > c^2\}$, то $\chi_{набл.}^2 > \chi_{1-\alpha}^2(n)$;

ако $H_1 = \{DX < c^2\}$, то $\chi_{набл.}^2 < \chi_{\alpha}^2(n)$.

- **При неизвестно EX : Статистика:** $\chi_{набл.}^2 = \frac{(n-1)\tilde{s}_x^2}{c^2} \sim \chi^2(n-1)$.

Критична област: ако $H_1 = \{DX \neq c^2\}$: $\chi_{набл.}^2 < \chi_{\frac{\alpha}{2}}^2(n-1)$, $\chi_{набл.}^2 > \chi_{1-\frac{\alpha}{2}}^2(n-1)$;

ако $H_1 = \{DX > c^2\}$: $\chi_{набл.}^2 > \chi_{1-\alpha}^2(n-1)$;

ако $H_1 = \{DX < c^2\}$: $\chi_{набл.}^2 < \chi_{\alpha}^2(n-1)$.

Тук с $\chi_{\alpha}^2(s)$ и $t_{\alpha}(s)$ са означени квантилите от ред α на χ^2 -разпределението и t -разпределението с s степени на свобода, които се взимат от таблиците на стр. 170 и 169.

Забележка 31.1. Освен с определяне на критичната област, проверката на хипотезата $H_0 = \{\theta = \theta_0\}$ може да се извърши и чрез така наречената p -стойност (p -value). Основните етапи тук са:

1. От извадката се изчислява наблюдаваната стойност $K_{набл.}$.
2. В зависимост от конкуриращата хипотеза::
 - При $H_1: \{\theta > \theta_0\}$ изчисляваме $p = P(K > K_{набл.})$
 - При $H_1: \{\theta < \theta_0\}$ изчисляваме $p = P(K < K_{набл.})$
 - При $H_1: \{\theta \neq \theta_0\}$ изчисляваме $p = P(K < K_{набл.}) + P(K > K_{набл.})$
3. **Взимане на решение.**
 - Ако $p < \alpha$, то $K_{набл.}$ се намира в критичната област и хипотезата H_0 се отхвърля.
 - Ако $p \geq \alpha$, то нямаме основание за отхвърляне на хипотезата H_0

Този начин за проверка на хипотезата е удобен при възможност за пресмятане на стойностите на функцията $F_K(x)$ на разпределение на статистическия критерий или при наличието на подробни таблици за разпределението на статистическия критерий.

Пример 31.1. От голяма партида резистори са избрани 36 резистора. Средната стойност на съпротивлението X на резисторите е $9,5k\Omega$. При ниво на значимост $\alpha=0,1$ да се провери хипотезата, че извадката е взета от партида, за която съпротивлението на резисторите е $10k\Omega$, при двустранна конкурираща хипотеза, ако: а) дисперсията на X е известна и равна на $4k\Omega^2$; б) дисперсията на X е неизвестна, а изборната поправена дисперсия е равна на $2,25k\Omega^2$;

Решение. По извадка, за която $\bar{x}=9,5$, проверяваме хипотезата $H_0=\{EX=10\}$ при конкурираща хипотеза $H_1=\{EX\neq 10\}$.

а) Средно квадратичното отклонение $\sigma X=\sigma=\sqrt{4}=2$ е известно. Тогава статистически критерий е $Z\sim N(0,1)$, а статистиката на критерия е $\frac{\bar{X}-a}{\sigma/\sqrt{n}}$.

1) Параметрите на статистиката се определят от обема $n=36$ и от нулевата хипотеза, според която $a=10$, $\sigma=2$, а наблюдаваната стойност на $\bar{X}$ е 9,5. Следователно, наблюдаваната стойност на критерия е

$$Z_{набл} = \frac{9,5-10}{2/\sqrt{36}} = \frac{-0,5}{1/3} = -1,5.$$

2) При ниво на значимост $\alpha=0,1$ определяме критичната стойност

$$Z = Z_{1-\frac{\alpha}{2}} = Z_{0,95} = 1,65.$$

3) Тъй като $|-1,5| < 1,65$, то хипотезата H_0 се приема.

б) Тук оценка на средно квадратичното отклонение е $\tilde{s}_x = \sqrt{2,25} = 1,5$. Като критерий използваме t -разпределението с

$n=36-1=35$ степени на свобода, $t_{набл} = \frac{9,5-10}{1,5/\sqrt{36}} = \frac{-3}{1,5} = -2$, критичната

стойност е $t_{1-\frac{\alpha}{2}}(35) = t_{0,95}(35) \approx 1,69$. Тъй като $|-2| > 1,69$, то хипотезата H_0 се отхвърля.

Пример 31.2. Да се намери мощността на критерия и вероятността за грешка от втори род β в пример 31.1а), ако е вярна конкуриращата хипотеза и действителната средна стойност на съпротивлението на резисторите е равно на $9k\Omega$.

Решение. В пример 31.1а) получихме, че критичната област при ниво на значимост 0,1 и областта на приемане са съответно $D=\{Z < -1,65 \cup Z > 1,65\}$ и $\bar{D}=\{Z \in [-1,65; 1,65]\}$. Статистиката на критерия

$$Z = \frac{\bar{X}-10}{2/\sqrt{36}} \quad (31.1)$$

изчислихме при предположение, че хипотезата $H_0=\{EX=10\}$ е вярна.

От (31.1) намираме съответната област на приемане за стойностите на величината $\bar{X}$ като решим относно $\bar{X}$ неравенството $-1,65 \leq \frac{\bar{X}-10}{1/3} \leq 1,65$, от което получаваме, че $\bar{X} \in [9,45; 10,55]$. Наблюдаваната стойност $\bar{x}=9,5$ е число в този интервал и затова нямаме основание да отхвърлим хипотезата H_0 .

Но в действителност е вярна конкуриращата хипотеза, по-конкретно, хипотезата $H_1=\{EX=9\}$, т.е. $\bar{X} \sim N(9; 1/3)$. Тогава намерената област $\bar{X} \in [9,45; 10,55]$ е област на грешка от втори род.

Първо ще намерим вероятността β за грешка от втори род $\beta = P(9,45 \leq \bar{X} \leq 10,55)$ или по формула (13.4):

$$\beta = F\left(\frac{10,55-9}{0,3333}\right) - F\left(\frac{9,45-9}{0,3333}\right) = F(4,65) - F(1,35) = 0,9999 - 0,9115 = 0,0884.$$

Следователно, мощността на критерия е $1-\beta = 1-0,0884 \approx 0,91$, т.е. при вярна хипотеза $H_1=\{EX=9\}$ и обем на извадката $n=36$ мощността на критерия е 0,91.

Пример 31.3. Точността на автомат, произвеждащ детайли, се характеризира с дисперсията на дължината X на детайлите. Ако тази величина е по-голяма от $4mm^2$, то автоматът трябва да се пренастрои. Извадъчната дисперсия на 15 случайно избрани детайла е равна на $6,8mm^2$. Необходимо ли е да се настройва автомата, ако приемем ниво на значимост а) 0,01; б) 0,1?

Решение. Приемаме, че дължината X на детайлите има нормално разпределение. Автоматът трябва да се настройва, ако $DX > 4$. Тъй като нулевата хипотеза трябва да бъде проста, то избираме $H_0=\{DX=4\}$ и $H_1=\{DX>4\}$ като хипотезата H_0 ще проверим при неизвестно математическо очакване. Според правилото за проверка на нулевата

$$\text{хипотеза изчисляваме } \chi^2_{набл} = \frac{(n-1)\tilde{s}_x^2}{c^2} = \frac{14\left(\frac{15}{14} \cdot 6,8\right)}{4} = 25,5.$$

Критичната стойност и решението за настройване на автомата определяме в зависимост от нивото на значимост:

а) $\chi^2_{кр} = \chi^2_{1-\alpha}(n-1) = \chi^2_{0,99}(14) = 29,14$. Тъй като $\chi^2_{набл} < \chi^2_{кр}$, то приемаме нулевата хипотеза, т.е. не е необходимо настройване на автомата.

б) $\chi^2_{кр} = \chi^2_{1-\alpha}(n-1) = \chi^2_{0,9}(14) = 21,06$. Тъй като $\chi^2_{набл} > \chi^2_{кр}$, то нулевата хипотеза се отхвърля и в този случай е необходима настройка на автомата.

Пример 31.4. По извадка с обем n и при дадено ниво на значимостта α да се проверят следните хипотези за нормално разпределената величина X като се използват p -стойности.

а) $H_0 = \{EX = 10\}$, $H_1 = \{EX \neq 10\}$, $n = 36$, $\alpha = 0,1$, ако $\bar{x} = 9,3$ и генералната дисперсия е $\sigma^2 = 4$;

б) $H_0 = \{EX = 10\}$, $H_1 = \{EX < 10\}$, $n = 36$, $\alpha = 0,1$, ако $\bar{x} = 9,5$ и изборната поправена дисперсия е $6,25k\Omega^2$ (виж пример 31.1а);

в) $H_0 = \{DX = 4\}$, $H_1 = \{DX > 4\}$, $n = 15$, $\alpha = 0,1$, ако EX е неизвестно.

Решение. а) При използване p -стойности извършваме следното:

$$1) \text{ Изчисляваме } Z_{\text{набл}} = \frac{9,5 - 10}{2/\sqrt{36}} = \frac{-0,5}{1/3} = -1,5.$$

2) За двустранна критична област изчисляваме

$$p = P(|Z| > 1,5) = 1 - P(|Z| \leq 1,5) = 1 - [2F(1,5) - 1] = 2 - 2 \cdot 0,9332 = 0,1336.$$

3) Тъй като $p > \alpha$, то наблюдаваната стойност се намира в областта на приемане, т.е. нямаме основание за отхвърляне на хипотезата H_0 .

$$б) \text{ Изчисляваме: } 1) t_{\text{набл}} = \frac{9,3 - 10}{2,5/\sqrt{36}} = \frac{-4,2}{2,5} = -1,68.$$

2) хипотезата $H_1 = \{EX < 10\}$ е лявостранна, следователно, за величината $T \sim t(n-1)$ (t -разпределение с $n-1 = 36-1 = 35$ степени на свобода трябва да изчислим вероятността $P(T < -1,68)$. Тъй като плътността на t -разпределението е четна функция, то

$$P(T < -1,68) = P(T > 1,68) = 1 - P(T < 1,68).$$

В таблицата за квантилите на t -разпределението не са дадени данни за разпределението с 35, нито има точно стойността 1,68, но може да използваме, че големи стойности на n t -разпределението е близко до $N(0,1)$, затова може да приемем, че $P(T < 1,68) > 0,95$. Тогава $p = P(T < -1,68) = 1 - P(T < 1,68) < 1 - 0,95 = 0,05$.

3) Следователно $p \approx 0,05 < \alpha$ и наблюдаваната стойност се намира в критичната област и нулевата хипотеза трябва да се отхвърли. (Ако разполагаме с таблица от критичните стойности на t -разпределението, то $P(T < -1,68) = P(T > 1,68) = 0,050932$.)

$$в) 1) \chi^2_{\text{набл}} = \frac{(n-1)\tilde{s}_x^2}{c^2} = \frac{14 \left(\frac{15}{14}, 6,8 \right)}{4} = 25,5,$$

2) Тъй като $H_1 = \{DX > 4\}$, то трябва да изчислим $p = P(\chi^2(14) > 25,5)$.

Ако разполагаме с подробна таблица на критичните точки на χ^2 (за 14 степени на свобода) определяме, че $p = 0,02938 > 0,01$, т.е. нямаме

основание да отхвърлим нулевата хипотеза.

Забележка. Ако използваме таблицата на стр. 170, то $p = 1 - P(\chi^2(14) < 25,5)$. Тъй като на реда за 14 степени числото 25,5 се намира между квантилите $\chi^2_{0,95}(14) = 23,68$ и $\chi^2_{0,975}(14) = 26,12$, то $0,95 < P(\chi^2_5(14) < 25,5) < 0,975$ и $p < 1 - 0,975 \approx 0,025$, откъдето отново заключаваме, че няма основание за отхвърлянето на H_0 . ♦

Упражнения (задачи върху §§30, 31).

1. Дадена е извадка 17, 15, 5, 9, 13, 42, 8, 24, 34, 38, 29, 6. Групирайте данните в интервали (0,10), (10,20), (20,30), (30,40) и (40,50). а) да се намерят извадъчните средна и средно квадратично отклонение на извадката; б) да се начертае хистограмата на относителните честоти. в) да се провери хипотезата H_0 , че величината X е разпределена равномерно в интервала [0, 50] (с ниво на значимост $\alpha = 0,05$).

2. При изпълнение на определени условия на експлоатация е установено, че средната продължителност на безотказна работа на голяма партида уреди е 1000 със средно квадратично отклонение 100 часа. Изборната средна на случайна извадка с обем 25 се е оказала 970 часа като средно квадратичното отклонение на извадката съвпада с това на цялата партида. Може ли да твърдим, че условията за експлоатация на цялата партида са спазени, ако приемем ниво на значимост: а) 0,10; б) 0,01?

3. Да се реши задача 2 при условие, че средно квадратичното отклонение на партидата е 120.

4. Твърди се, че изработваните в завод детайли имат диаметър 10 мм. Използвайки едностранен критерий да се провери с ниво на значимост 0,05 това твърдение, ако за извадка с обем $n = 16$ се е оказало, че средния диаметър е равен на 10,3, считайки, че а) дисперсията е известна и е равна на 1 мм^2 ; б) поправената дисперсия на извадката е $1,21 \text{ мм}^2$.

5. Изпробва се нов метод за измерване на разстояния. Направени са 10 измервания на един и същ еталон и е изчислено, че поправеното средно квадратично отклонение от измерванията е 100 мм^2 . Съгласува ли се този резултат с хипотезата, че дисперсията на грешките от измерванията по този метод не са по-големи от 50 мм^2 ?

6. От генерална съвкупност е извлечена извадка с обем 150:

x_i	6	9	12	15	18	21	24	27	30	33
m_i	1	4	13	23	22	29	29	16	11	2

Да се провери с 10%-но ниво на значимост хипотезата за нормално разпределение на признака X на генералната съвкупност.

7. При изпитание на апаратура се отчита броя X на повредите. Резултатите от 59 изпитания са представени в таблицата: $\frac{x_i}{m_i} \begin{vmatrix} 0 & 1 & 2 & 3 \\ 42 & 1 & 4 & 3 \end{vmatrix}$. Да се провери хипотезата

$X \sim Po(\lambda)$ с ниво на значимост а) $\alpha = 0,01$; б) $\alpha = 0,05$;

8. Дадена е извадка (X - брой на децата в 25 семейства) 3, 2, 0, 1, 4, 5, 3, 2, 4, 1, 0, 2, 5, 5, 2, 1, 2, 1, 0, 0, 3, 6, 2, 1, 0. а) да се начертае полигонът на относителната честота. Изказано е предположение, че 20% от семействата в даден район нямат деца, 30% от семействата имат едно, 30% - две деца, а останалите 20% - повече от две деца. Да се провери това предположение на базата на дадената извадка с ниво на значимост 0,05.

ИЗСЛЕДВАНЕ НА ДВЕ ИЗВАДКИ

§32. Зависими и независими извадки. Изследване на две зависимы извадки.

В много случаи с помощта на статистически методи се търси оценка на различията между две отделни групи от обществото, на ефекта от прилагането на даден метод за лечение или технологично нововъведение. Например, въздействието на дадено лекарство се изследва с два подхода:

- 1) Взимат се две отделни съвкупности от доброволци като на едните се дава новото лекарство, а на другите се дава безобидно лекарство, без истински ефект, която роля е само психологическа.
- 2) Изследва се една и съща съвкупност от доброволци, като се отчитат резултатите преди и след приемане на лекарството.

Двата подхода се различават главно по избора на обектите на наблюденията: в първия случай наблюденията се извършват върху две случайни извадки, които нямат общ елемент помежду си. Извадките могат да бъдат от една или от две различни генерални съвкупности и да имат различен обем. Такива извадки се наричат независими и те ще бъдат разгледани в §34.

При втория подход наблюденията се извършват върху елементите от една и съща съвкупност или върху елементите на зависимы съвкупности. Такива извадки се наричат зависими. Получените съвкупности от наблюдавани стойности $(x_1, \dots, x_n)$ и $(y_1, \dots, y_n)$ се разглеждат като наблюдавани стойности на различни признаци X и Y на генералната съвкупност.

Зависимите извадки имат еднакъв обем и се изследват като се съпоставят резултатите от наблюденията x_i и y_i на i -тия обект, при изучаваните условия (напр. преди и след прилагането на лекарството). За целта се разглеждат разликите $d_i = x_i - y_i$, чрез които се получава се нова съвкупност от данни. Прилагайки различни статистически методи за данните $(d_1, \dots, d_n)$ – точкови оценки, доверителни интервали и хипотези за средната $\bar{d}$ или за дисперсията $\tilde{s}_d^2$, се правят оценки и изводи за влиянието на новите условия върху изучаваната съвкупност от обекти, а от там и на цялата генерална съвкупност.

Ще разгледаме случая, когато X и Y са **нормално разпределени** с математически очаквания $EX = \mu_x$ и $EY = \mu_y$ и дисперсии $DX = \sigma_x^2$ и $DY = \sigma_y^2$. Разликата $D = X - Y$ също има нормално разпределение и затова може да приложим методите за изследване на нормално разпределени признаци.

Пример 32.1. Изучава се влиянието на лекарство върху диастоличното кръвно налягане. Измерено е диастоличното кръвно

налягане на 8 пациента преди и след приемането на лекарството. Получени се следните данни

x_i	93	106	87	102	95	88	110	92
y_i	92	102	89	101	96	88	105	92
$d_i = x_i - y_i$	1	4	-2	1	-1	0	5	0

а) да се намери доверителен интервал за средното изменение на диастоличното кръвно налягане с доверителна вероятност $\gamma = 0,99$;

б) Може ли да се твърди на базата на наличните данни, че лекарството не влияе на диастоличното кръвно налягане, ако приемем ниво на значимост $\alpha = 0,01$?

Решение. Изчисляваме: $\bar{d} = \frac{1}{8}(1+4-2+1-1+0+5+0) = 1$.

$$\tilde{s}_d^2 = \frac{1}{n-1} \sum (d_i - \bar{d})^2 = 5,71, \quad \tilde{s}_d = \sqrt{\tilde{s}_d^2} = \sqrt{5,71} = 2,39.$$

а) $\bar{d} = 1$ е точкова оценка за разликата $\mu_x - \mu_y$, а доверителният интервал за тази разлика ще намерим по формула (28.2):

1) намираме квантила от ред $(1+\gamma)/2 = 0,995$ на t -разпределението с $n-1 = 8-1 = 7$ степени на свобода: $t_{0,995}(7) = 3,50$.

2) Представителната грешка е $\delta = t_{0,995}(7) \frac{\tilde{s}_d}{\sqrt{n}} = 3,50 \frac{2,39}{\sqrt{8}} = 2,957$.

3) Следователно, с вероятност 0,99 разликата $\mu_x - \mu_y$ е число от интервала $(1-2,957; 1+2,957) = (-1,957; 3,957)$;

б) Проверяваме хипотезата $H_0 = \{\mu_x = \mu_y\}$, т.е. $\mu_x - \mu_y = 0$, при ниво на значимост 0,01. Ако конкуриращата хипотеза е $H_1 = \{\mu_x \neq \mu_y\}$, то можем да използваме получения в а) резултат. Наистина, тъй като $0 \in (-1,957; 3,957)$ с вероятност 0,99, то вероятността да приемем вярна хипотеза H_0 е 0,99, а вероятността да отхвърлим вярна хипотеза H_0 ще бъде $1-0,99 = 0,01$, което съвпада с нивото на значимост. Следователно, може да приемем с ниво на значимост 0,01, че лекарството не влияе на диастоличното кръвно налягане (до същия извод се достига, и ако се приложи проверка на хипотезата H_0).

Пример 32.2. Изследване, базирано на две зависимы извадки с обем 17, има за цел да отхвърли или приеме нулевата хипотеза за равенство на генералните средни ($H_0 = \{\mu_x = \mu_y\}$) при конкурираща хипотеза $H_1 = \{\mu_x > \mu_y\}$. Намерено е, че средната на разликите d_i е $\bar{d} = 5,7$, а поправеното средно квадратично отклонение на разликите е

$\tilde{s}_d = 4,8$. Достатъчни ли са наличните данни за отхвърляне на нулевата хипотеза при ниво на значимост 0,05?

• **Решение.** Ще използваме теста за проверка на хипотезата $H_0 = \{\mu_d = 0\}$ при конкурираща хипотеза $H_1 = \{\mu_d > 0\}$.

• 1) Наблюдаваната стойност е $t_{набл.} = \frac{\bar{d} - 0}{\tilde{s}_d / \sqrt{n}} = \frac{5,7 \cdot \sqrt{17}}{4,8} = 4,896$.

• 2) Критичната област се определя по формула (31.2), съгласно която изчисляваме $t_{кр.} = t_{1-\alpha}(n-1) = t_{0,95}(16) = 1,75$. Тъй като $t_{набл.} > t_{кр.}$, то данните са достатъчни за отхвърляне на нулевата хипотеза.

• **Упражнение.**

• 1. Сравнява се времето, необходимо за изпълнение на програми, написани на два различни програмни езика. За 12 програми времето (в минути) за

изпълнение е следното:

x_i	17	16	21	14	18	24	13	23	14	16	18	21
y_i	18	14	19	11	23	21	15	24	13	10	20	19

Да се

намери доверителен интервал за средната разлика $d_i = x_i - y_i$ с доверителна вероятност $\gamma = 0,95$. Може ли да се направи извод за по-голяма ефективност на един от двата езика?

§33. Критерии за сравняване на дисперсиите и математическите очаквания на две нормално разпределени генерални съвкупности.

Нека за изследване на зависимостите между признаците X и Y на две генерални съвкупности са дадени независимите извадки $(X_1, \dots, X_{n_x})$ с обем n_x от първата генерална съвкупност и $(Y_1, \dots, Y_{n_y})$ с обем n_y на втората генерална съвкупност. Сравняването на признаците X и Y се извършва чрез:

• Диаграмите тип бокс-плот на двете извадки за сравняване на общите закономерности (виж пример 33.1)

• Разликата на извадъчните средни $\bar{x} - \bar{y}$ за сравняване на генералните средни $EX = \mu_x$ и $EY = \mu_y$

• Отношенията на дисперсиите им $\tilde{s}_x^2$ и $\tilde{s}_y^2$ за сравняване на дисперсиите $DX = \sigma_x^2$ и $DY = \sigma_y^2$

1. Сравняване на дисперсии. Нека за изследване на величините $X \sim N(\mu_x, \sigma_x^2)$ и $Y \sim N(\mu_y, \sigma_y^2)$ разполагаме с две независими извадки като с X е означена тази величина, за която $\tilde{s}_x^2 > \tilde{s}_y^2$. За съпоставяне на

дисперсиите на величините се използва, че отношението $\frac{\tilde{s}_x^2 / \sigma_x^2}{\tilde{s}_y^2 / \sigma_y^2}$ има

разпределение на Фишер и се извършва по следния начин:

Намиране на доверителни интервали за отношението $\frac{\sigma_x^2}{\sigma_y^2}$ с

доверителна вероятност γ :

1) μ_x и μ_y са известни: $\frac{\sigma_x^2}{\sigma_y^2} \in \left(\frac{s_{0x}^2 / s_{0y}^2}{F_1}, \frac{s_{0x}^2 / s_{0y}^2}{F_2} \right)$, където $s_{0x}^2 > s_{0y}^2$ са:

$$s_{0x}^2 = \frac{1}{n_x - 1} \sum (x_i - \mu_x)^2, s_{0y}^2 = \frac{1}{n_y - 1} \sum (y_i - \mu_y)^2 \quad (33.1)$$

$$F_1 = F_{\frac{1+\gamma}{2}}(n_x, n_y), F_2 = F_{\frac{1-\gamma}{2}}(n_x, n_y) - \text{квантили на } F\text{-разпределението с}$$

степени на свобода n_x и n_y .

2) μ_x и μ_y са неизвестни: $\frac{\sigma_x^2}{\sigma_y^2} \in \left(\frac{\tilde{s}_x^2 / \tilde{s}_y^2}{F_1}, \frac{\tilde{s}_x^2 / \tilde{s}_y^2}{F_2} \right)$, ($\tilde{s}_x^2 > \tilde{s}_y^2$) (33.2)

където: $F_1 = F_{\frac{1+\gamma}{2}}(n_x - 1, n_y - 1)$, $F_2 = F_{\frac{1-\gamma}{2}}(n_x - 1, n_y - 1)$ са квантили на F -разпределението с степени на свобода $n_x - 1$ и $n_y - 1$.

Забележка 33.1. Обикновено са съставени таблици за критичните точки на F -разпределението, като квантилите и критичните точки са свързани по формули (14.2).

Пример 33.1. Изследва се диаметърът на детайли, произведени от две машини. От произведените детайли са получени две извадки с обеми $n_x = 25$ и $n_y = 13$ и са изчислени, $\tilde{s}_x^2 = 0,35$, $\tilde{s}_y^2 = 0,40$. Да се намери 90% доверителен интервал за отношението σ_y^2 / σ_x^2 .

Решение. 1) Изчисляваме $\frac{\tilde{s}_y^2}{\tilde{s}_x^2} = \frac{0,40}{0,35} = 1,1428$ (тук $\tilde{s}_y^2 > \tilde{s}_x^2$)

2) На стр.170 са дадени **критичните точки** на F -разпределението за ниво на значимост 0,05, затова (виж формули 14.2):

$$F_1 = F_{\frac{1+\gamma}{2}}(n_y - 1, n_x - 1) = F_{1,9}(12, 24) = F_{0,95}(12, 24) = F_{0,05}^{kp}(12, 24) = 2,5055$$

$$F_2 = F_{\frac{1-\gamma}{2}}(n_y - 1, n_x - 1) = F_{0,05}(12, 24) = \frac{1}{F_{0,05}^{kp}(24, 12)} = \frac{1}{2,1834} = 0,458$$

По формула (33.2) получаваме $\frac{\sigma_y^2}{\sigma_x^2} \in \left(\frac{1,1428}{2,5055}, \frac{1,1428}{0,458} \right) = (0,456; 2,495) \blacklozenge$

В зависимост от това дали μ_x и μ_y са известни или не, се определят следните правила за проверка на хипотезата за равенство на дисперсиите на две независими извадки:

Проверка на хипотезата $H_0 = \{\sigma_x^2 = \sigma_y^2\}$ при ниво на значимост α .

1) μ_x и μ_y са известни:

- Наблюдавана стойност: $F_{набл.} = \frac{S_{0x}^2}{S_{0y}^2}$ като $s_{0x}^2 > s_{0y}^2$ (формула (33.1)).
- Критична област:
 $F_{набл.} > F_{кр.}$ при $H_1 = \{\sigma_x^2 > \sigma_y^2\}$, където $F_{кр.} = F_{1-\alpha}(n_x, n_y)$.
 $F_{набл.} > F_{кр.}$ при $H_1 = \{\sigma_x^2 \neq \sigma_y^2\}$, където $F_{кр.} = F_{1-\frac{\alpha}{2}}(n_x, n_y)$.

2) μ_x и μ_y са неизвестни:

- Наблюдавана стойност: $F_{набл.} = \frac{\tilde{S}_x^2}{\tilde{S}_y^2}$ като $(\tilde{S}_x^2 > \tilde{S}_y^2)$.
- Критична област:
 $F_{набл.} > F_{кр.}$ при $H_1 = \{\sigma_x^2 > \sigma_y^2\}$, където $F_{кр.} = F_{1-\alpha}(n_x-1, n_y-1)$
 $F_{набл.} > F_{кр.}$ при $H_1 = \{\sigma_x^2 \neq \sigma_y^2\}$, където $F_{кр.} = F_{1-\frac{\alpha}{2}}(n_x-1, n_y-1)$

За изследване на разликата $\mu_x - \mu_y$ за нормално разпределени величини $X \sim N(\mu_x, \sigma_x)$ и $Y \sim N(\mu_y, \sigma_y)$, се използва t -разпределението по следния начин:

Доверителни интервали за разликата $\mu_x - \mu_y$:

1) σ_x^2 и σ_y^2 -известни:

$$\mu_x - \mu_y \in \left(\bar{x} - \bar{y} - Z_{\frac{1+\gamma}{2}} \sigma, \bar{x} - \bar{y} + Z_{\frac{1+\gamma}{2}} \sigma \right), \quad \text{където } \sigma = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}.$$

2) $\sigma_x^2 = \sigma_y^2$, но не са известни:

$$\mu_x - \mu_y \in \left(\bar{x} - \bar{y} - t \sqrt{s^2 \left(\frac{1}{n_x} + \frac{1}{n_y} \right)}, \bar{x} - \bar{y} + t \sqrt{s^2 \left(\frac{1}{n_x} + \frac{1}{n_y} \right)} \right),$$

$$\text{където } s^2 = \frac{(n_x-1)\tilde{S}_x^2 + (n_y-1)\tilde{S}_y^2}{n_x + n_y - 2}, \quad t = t_{\frac{\gamma+1}{2}}(n_x + n_y - 2)$$

$$3) \sigma_x^2 \neq \sigma_y^2 \text{ -неизвестни: } \mu_x - \mu_y \in \left(\bar{x} - \bar{y} - t \sqrt{\frac{\tilde{S}_x^2}{n_x} + \frac{\tilde{S}_y^2}{n_y}}, \bar{x} - \bar{y} + t \sqrt{\frac{\tilde{S}_x^2}{n_x} + \frac{\tilde{S}_y^2}{n_y}} \right),$$

$$\text{където } t = t_{\frac{\gamma+1}{2}}(k^*), \quad k^* \text{ е цялата част на израза } \frac{\left[(\tilde{S}_x^2/n_x) + (\tilde{S}_y^2/n_y) \right]^2}{\frac{(\tilde{S}_x^2/n_x)^2}{n_x-1} + \frac{(\tilde{S}_y^2/n_y)^2}{n_y-1}}.$$

Проверка на хипотезата $H_1 = \{\mu_x = \mu_y\}$ при ниво на значимост α :

1) σ_x^2 и σ_y^2 -известни:

- Наблюдавана стойност: $Z_{набл.} = \frac{\bar{x} - \bar{y}}{\sigma}$, $\sigma = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$
- Критична област:
 $Z_{набл.} > Z_{кр.}$ при $H_1 = \{\mu_x > \mu_y\}$, където $Z_{кр.} = Z_{1-\alpha}$
 $Z_{набл.} < Z_{кр.}$ при $H_1 = \{\mu_x < \mu_y\}$, където $Z_{кр.} = -Z_{1-\alpha}$
 $|Z_{набл.}| > Z_{кр.}$ при $H_1 = \{\mu_x \neq \mu_y\}$, където $Z_{кр.} = Z_{1-\frac{\alpha}{2}}$

2) $\sigma_x^2 = \sigma_y^2$ - неизвестни и хипотезата $H'_0 = \{\sigma_x^2 = \sigma_y^2\}$ се приема:

- Наблюдавана стойност:

$$t_{набл.} = \frac{\bar{x} - \bar{y}}{\sqrt{s^2 \left(\frac{1}{n_x} + \frac{1}{n_y} \right)}}, \quad s^2 = \frac{(n_x-1)\tilde{S}_x^2 + (n_y-1)\tilde{S}_y^2}{n_x + n_y - 2} \quad (33.2)$$
- Критична област:
 $t_{набл.} > t_{кр.}$ при $H_1 = \{\mu_x > \mu_y\}$, където $t_{кр.} = t_{1-\alpha}(n_x + n_y - 2)$
 $t_{набл.} < t_{кр.}$ при $H_1 = \{\mu_x < \mu_y\}$, където $t_{кр.} = -t_{1-\alpha}(n_x + n_y - 2)$
 $|t_{набл.}| > t_{кр.}$ при $H_1 = \{\mu_x \neq \mu_y\}$, където $t = t_{\frac{\gamma+1}{2}}(n_x + n_y - 2)$

3) $\sigma_x^2 \neq \sigma_y^2$ неизвестни и хипотезата $H'_0 = \{\sigma_x^2 = \sigma_y^2\}$ се отхвърля:

- Наблюдавана стойност: $t_{набл.} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\tilde{S}_x^2}{n_x} + \frac{\tilde{S}_y^2}{n_y}}}$
- Критична област:
 $t_{набл.} > t_{кр.}$ при $H_1 = \{\mu_x > \mu_y\}$, където $t_{кр.} = t_{1-\alpha}(k^*)$
 $t_{набл.} < t_{кр.}$ при $H_1 = \{\mu_x < \mu_y\}$, където $t_{кр.} = -t_{1-\alpha}(k^*)$
 $|t_{набл.}| > t_{кр.}$ при $H_1 = \{\mu_x \neq \mu_y\}$, където $t_{кр.} = t_{1-\frac{\alpha}{2}}(k^*)$

Тук k^* е цялата част на израза $\frac{\left[(\tilde{S}_x^2/n_x) + (\tilde{S}_y^2/n_y) \right]^2}{\frac{(\tilde{S}_x^2/n_x)^2}{n_x-1} + \frac{(\tilde{S}_y^2/n_y)^2}{n_y-1}}.$

Забележка 33.2. Напомняме отново: тъй като има таблици за критичните точки на F -разпределението за стойности, по-

големи от единица, то приемаме, че в горните формули с X е означена тази величина, за която $\tilde{s}_x^2 > \tilde{s}_y^2$.

На практика сравняване на параметрите на две независими извадки се налага при сравняване на два метода на работа, на две технологии и т.н.

Пример 32.1. Предложена е нова технология за производство на лампи; за която се твърди, че осигурява по-голямо светлоотдаване и по-малко отклонение от средното светлоотдаване. За проверка на хипотезите са проверени 8 лампи, произведени по старата технология, стойностите на светлоотдаване на които са (в лм/вт):

62, 73, 80, 79, 63, 77, 81, 75

и 9 лампи по новата технология, със светлоотдаване:

75, 78, 77, 68, 73, 79, 72, 71, 86.

а) да се сравнят диаграмите от тип бокс-плот.

б) Да се проверят хипотезите с ниво на значимост $\alpha=0,01$.

Решение.

а) Намираме вариационните редове:

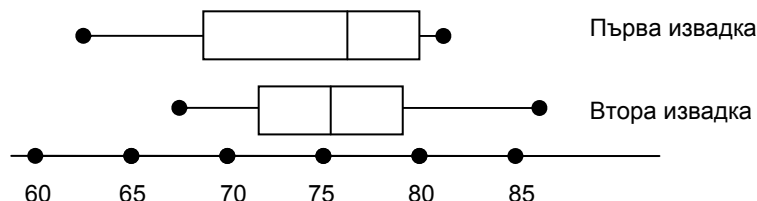
Първа извадка: 62, 63, 73, 75, 77, 79, 80, 81,

откъдето $Q_1 = \frac{63+73}{2} = 68$, $Q_2 = \frac{75+77}{2} = 76$, $Q_3 = \frac{80+81}{2} = 80,5$.

Втора извадка: 68, 71, 72, 73, 75, 77, 78, 79, 86,

откъдето: $Q'_1 = 72$, $Q'_2 = 75$, $Q'_3 = 78$.

Начертаваме диаграмите от тип бокс-плот:



Изводи: Медианите (които са оценки за средната стойност на величините) са близки $\Rightarrow$ средните стойности не се различават съществено.

$Q'_3 - Q'_1 = 78 - 72 = 6$, $Q_3 - Q_1 = 80,5 - 68 = 12,5 \Rightarrow$ отклоненията от средните стойности на втората извадка са по-малки от тези на първата.

б) Нека X е светлоотдаването на лампа, произведена по старата технология. Тогава EX е средното светлоотдаване на лампите, а σX е отклонението от средното светлоотдаване. Аналогично, ако Y е светлоотдаването на лампа, произведена по новата технология, то EY и σY са съответно средното светлоотдаване и разсейването около него на произведените по втората технология лампи.

От извадките изчисляваме оценките на EX , σX , EY и σY :

$$\bar{x} = 73,75, \quad \tilde{s}_x^2 = 55,07; \quad \bar{y} = 75,44, \quad \tilde{s}_y^2 = 28,73.$$

Съгласно условието на задачата, издигнати са хипотезите: при производството по втората технология имаме по-добро светлоотдаване т.е. $EY > EX$ и по-малко разсейване, т.е. $\sigma^2 Y < \sigma^2 X$. Тъй като нулевата хипотеза трябва да е проста, то при ниво на значимост $\alpha=0,01$ ще проверим хипотезите

$$H_0 = \{\sigma_x^2 = \sigma_y^2\} \text{ срещу } H_1 = \{\sigma_x^2 > \sigma_y^2\} \text{ и}$$

$$H_0 = (EX = EY) \text{ срещу } H_1 = \{EX < EY\}$$

• Проверка на хипотезата $H_0 = \{\sigma_x^2 = \sigma_y^2\}$ срещу $H_1 = \{\sigma_x^2 > \sigma_y^2\}$ и ниво на значимост $\alpha=0,01$:

1) Наблюдаваната стойност е

$$F_{\text{набл.}} = \frac{\text{по-голяма дисперсия}}{\text{по-малка дисперсия}} = \frac{\tilde{s}_x^2}{\tilde{s}_y^2} = \frac{55,07}{28,73} = 1,916.$$

2) При дадената конкурираща хипотеза критичната стойност определяме като критична точка от ред $\alpha=0,01$ на разпределението

$$F(n_x - 1, n_y - 1) = F(7, 8), \text{ т.е. } F_{\text{кр.}} = F_{0,01}^{(\text{кр})}(7, 8) = 6,18.$$

3) Тъй като $F_{\text{набл.}} < F_{\text{кр.}}$, то хипотезата за равенство на дисперсиите се приема.

• Проверка на хипотезата $H_0 = (EX = EY)$ срещу $H_1 = \{EX < EY\}$.

1) Тъй като се установи, че $H_0 = \{\sigma_x^2 = \sigma_y^2\}$ не може да се отхвърли на базата на наличните данни, то наблюдаваната стойност на критерия изчисляваме по формула (33.2). Изчисляваме обединената дисперсия

$$s^2 = \frac{(n_x - 1)\tilde{s}_x^2 + (n_y - 1)\tilde{s}_y^2}{n_x + n_y - 2} = \frac{7 \cdot 55,07 + 8 \cdot 28,73}{8 + 9 - 2} = 41,022,$$

откъдето

$$t_{\text{набл.}} = \frac{\bar{x} - \bar{y}}{\sqrt{s^2 \left(\frac{1}{n_x} + \frac{1}{n_y} \right)}} = \frac{73,75 - 75,77}{\sqrt{41,022 \left(\frac{1}{8} + \frac{1}{9} \right)}} = -0,5444.$$

2) За критичната област определяме съгласно конкуриращата хипотеза като лявостранната област $D = \{t_{\text{набл.}} < t_{\text{кр.}}\}$, т.е. за която

$$P(t_{\text{набл.}} < t_{\text{кр.}}) = \alpha, \text{ получаваме: } t_{\text{кр.}} = -t_{1-\alpha}(n_x + n_y - 2) = t_{0,99}(15) = -2,60.$$

3) Така получаваме, че $t_{\text{набл.}} > t_{\text{кр.}}$ и хипотезата за равенство на средните стойности също се потвърждава.

Следователно, на основа на наличните данни не може да се приеме, че новата технология ще подобри параметрите на произвежданите лампи. ♦

Упражнения:

1. При измерване на физична величина X са получени резултатите -2, 0, -1, 1, 3.
 - а) да се намерят оценките за математическото очакване и дисперсията на величината X .
 - б) да се определи с надеждност $\gamma=0,95$ доверителният интервал за математическото очакване на величината X .
 - в) По друг метод са направени 6 измервания и е получено, че поправеното средно квадратично отклонение е $\tilde{s}_Y^2 = 5$. Да се провери с ниво на значимост $\alpha=0,02$ хипотезата $H_0 = \{DY = DX\}$ при конкурираща хипотеза $H_1 = \{DY \neq DX\}$.
2. От продукцията на две автоматични линии са взети извадки от $n_1=16$ и $n_2=12$ детали и са изчислени $\bar{x}_1=180$ мм и $\bar{x}_2=186$ мм. От предварителен анализ е установено, че неточността на изработката им са нормално разпределени величини с дисперсии съответно $\sigma_1^2=6$ мм² и $\sigma_2^2=11$ мм². С ниво на значимост $\alpha = 0.025$ да се провери хипотезата $H_0: \mu_1 = \mu_2$ срещу хипотезата $H_1: \mu_1 < \mu_2$.
3. Две съвкупности от обекти, на които се изучава количественият признак X с нормално разпределение, са подложени на контрол. От първото множество е направена извадка с обем $n_1=21$ и пресметната поправената дисперсия, която е равна на 0,75. От второто множество е направена извадка с обем $n_2=11$ с поправената дисперсия 0,25. Проверете хипотезата $H_0: DX_1=DX_2$, при конкурираща хипотеза $H_1: DX_1 > DX_2$ и ниво на значимост 0,05.
4. Предприятие има два автомата за производство на изделия, диаметърът на които се изследва. Взети са по 100 изделия за проверка. Средния диаметър на изработените от първия автомат изделия е 10,3, а стандартната грешка е 1,2, а за изработените от втория – съответно 9,8 и 1,6. С ниво на значимост 0,01 да се проверят хипотезите:
 - а) Точността на изработка на двата автомата е еднаква.
 - б) Двата автомата произвеждат изделия с еднакви диаметри.
5. При измерване на една и съща величина X са получени следните резултати $x_1=-4, x_2=0, x_3=-2, x_4=2, x_5=6$.
 - а) Да се намерят $\bar{x}$ и $\tilde{s}_X$.
 - б) Да се определи доверителният интервал за EX на случайната величина X с доверителна вероятност $\gamma=0,95$.
 - в) По друг метод от извадка с обем $n=9$ е получена извадъчна дисперсия $\tilde{s}_Y^2=16$. Да се провери с ниво на значимост $\alpha=0,02$ хипотезата за еднаква точност на двата метода, т.е. хипотезата $H_0=\{DX=DY\}$. Като конкурираща хипотеза да се разгледа $H_1=\{DX \neq DY\}$ (предполага се, че X има нормално разпределение).

КОРЕЛАЦИОНЕН И РЕГРЕСИОНЕН АНАЛИЗ

§34. Двумерни извадки. Числени характеристики на извадката и точкови оценки. Корелационна таблица.

Много често в практиката се разглеждат два признака X и Y на генералната съвкупност като целта тук е не само изучаване на всеки от признаците, но и намиране на зависимост между тях, ако съществува такава. Например, изучават се теглото и ръстът на населението, броят на пушачите и броят на заболяванията на дихателните пътища и т.н. В такъв случай в резултат на наблюденията се получават двойки числа (x_i, y_i) , $i=1, \dots, n$, затова получената извадка се нарича двумерна извадка с обем n . Тъй като приемаме, че обектите на изследване са случайно избрани, то ще считаме, че (x_i, y_i) са реализации на двумерната случайна величина (X, Y) .

Резултатите от наблюденията се подреждат във вид на таблица:

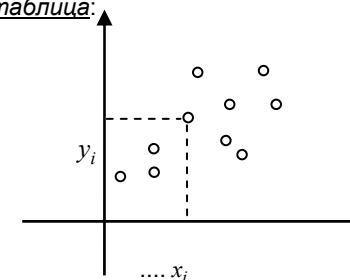
$$\begin{array}{c|cccc} x_i & x_1 & x_2 & \dots & x_n \\ y_i & y_1 & y_2 & \dots & y_n \end{array} \quad (34.1)$$

Графическо представяне на извадката е диаграмата на разсейването – съвкупност от точки с координати (x_i, y_i) в равнината Oxy (фиг. 1).

1. Групиране на данните. Данните, представени чрез таблица (34.1) са негрупирани.

Ако има повторения на наблюдаваните стойности по-удобно е представянето им в корелационна таблица:

x_i	x_1	$\dots$	x_k	m_{y_j}
y_j	m_{11}	$\dots$	m_{k1}	m_{y_1}
y_s	m_{1s}	$\dots$	m_{ks}	m_{y_s}
m_{x_i}	m_{x_1}	$\dots$	m_{x_k}	n



Фиг.34.1.

в която m_{ij} – брой на наблюденията (x_i, y_j) и $n = \sum_{i=1}^k \sum_{j=1}^s m_{ij}$ – обем на извадката. Последната колона се състои от сумите по съответния ред, $m_{y_j} = \sum_{i=1}^k m_{ij}$ – честотата на вариантата y_j , а последния ред – сумите по съответната колона, $m_{x_i} = \sum_{j=1}^s m_{ij}$ – честотата на вариантата x_i .

Пример 34.1. Дадена е извадка с обем $n=5$: (8,1), (10,3), (5,1), (8,2) и (9,3). Да се получат негрупираното и групираното разпределение на извадката (корелационна таблица).

Решение: Негрупирано представяне на данните: $\begin{matrix} x_i & 8 & 10 & 5 & 8 & 9 \\ y_i & 1 & 3 & 1 & 2 & 3 \end{matrix}$.

Корелационна таблица:

x_i	5	8	9	10	m_{y_j}
y_j	1	1	0	0	2
2	0	1	0	0	1
3	0	0	1	1	2
m_{x_i}	1	2	1	1	5

Възможно е групиране на данните по интервали:

(x_{i-1}, x_i)	(x_0, x_1)	...	(x_{k-1}, x_k)	
(y_{j-1}, y_j)	m_{11}	...	m_{k1}	m_{y_1}
(y_0, y_1)	$\vdots$			$\vdots$
(y_{s-1}, y_s)	m_{1s}	...	m_{ks}	m_{y_s}
	m_{x_1}	...	m_{x_k}	n

2. Числени характеристики.

1. Числени характеристики на съставните компоненти X и Y .

Очевидно, може разгледаме поотделно и да пресметнем всички познати ни числени характеристики на компонентите X и Y :

• За признака X : $\bar{x}$, s_x^2 , $\tilde{s}_x^2$, начални $\overline{x^p}$ и централни $s_x^{(p)}$ моменти, които са и точкови оценки на съответните моменти на величината X . Формулите за пресмятането им се получават от статистическото разпределение на X , например,

$$\text{извадъчна средна на } X: \bar{x} = \begin{cases} \frac{1}{n} \sum_{i=1}^n x_i, & \text{за негрупирани данни} \\ \frac{1}{n} \sum_{i=1}^k m_{x_i} x_i & \text{за групирани данни} \end{cases};$$

начален момент от ред p :

$$\bar{x}^p = \begin{cases} \frac{1}{n} \sum_{i=1}^n x_i^p, & \text{за негрупирани} \\ \frac{1}{n} \sum_{i=1}^k m_{x_i} x_i^p = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^s m_{ij} x_i^p & \text{за групирани} \end{cases};$$

извадъчна и поправена дисперсия:

$$s_x^2 = \frac{1}{n} \sum_{i=1}^k m_{x_i} (x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2 \text{ и } \tilde{s}_x^2 = \frac{n}{n-1} s_x^2.$$

• За признака Y : Аналогично получаваме $\bar{y}$, s_y^2 , $\tilde{s}_y^2$, начални $\overline{y^p}$ и централни $s_y^{(p)}$ моменти, които са и съответни точкови оценки на съответните моменти на величината Y .

2. Числени характеристики, отразяващи връзките между наблюдаваните стойности:

• Извадъчна ковариация:

$$s_{xy} = \begin{cases} \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) & \text{за негрупирани данни} \\ \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^s m_{ij} (x_i - \bar{x})(y_j - \bar{y}) & \text{за групирани данни} \end{cases}.$$

Извадъчната ковариация се явява отместена оценка на ковариацията $\text{cov}(X, Y)$ на двумерната случайна величина (X, Y) .

По-удобна формула за пресмятане на извадъчната ковариация:

$$s_{xy} = \overline{xy} - \bar{x} \cdot \bar{y}, \text{ където } \overline{xy} = \begin{cases} \frac{1}{n} \sum_{i=1}^n x_i y_i & \text{за негрупирани данни} \\ \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^s m_{ij} x_i y_j & \text{за групирани данни} \end{cases}.$$

• Поправена извадъчна корелация. $\tilde{s}_{xy} = \frac{n}{n-1} s_{xy}$ - поправена оценка на $\text{cov}(X, Y)$:

• Извадъчен коефициент на корелация (коефициент на Пирсън):

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} \quad (34.2)$$

Този коефициент е оценка на коефициента на корелация

$$\rho_{xy} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \text{ на случайната величина } (X, Y) \text{ и има свойствата:}$$

- 1) не зависи от мерните единици на съставните;
- 2) $-1 \leq r_{xy} \leq 1$;
- 3) $r = \pm 1 \Leftrightarrow$ между X и Y има линейна зависимост $aX + bY + c = 0$.
- 4) измерва силата на линейната зависимост между X и Y :

Силата на линейната корелация се определя по скалата:

- Ако $|r_{xy}| \leq 0,3$ – слаба зависимост (фиг. 35.1а);
- Ако $0,3 < |r_{xy}| < 0,8$ – средна зависимост (фиг. 35.1б);
- Ако $|r_{xy}| \geq 0,8$ – силна зависимост (фиг. 35.1в).

Рангов коефициент на корелация (коефициент на Спирмън).

Ранговият коефициент на корелация оценява силата на така наречената рангова корелационна зависимост между компонентите X и Y (зависимост). За определянето му от дадена една извадка (x_i, y_i) , $i=1, \dots, n$ подреждаме получените стойности x_i на X по големина във възходящ ред (т.е. получаваме вариационния ред $\hat{x}_i$).

Поредният номер x'_i на x_i във вариационния ред се нарича ранг на варианта.

По същия начин получаваме и ранговете y'_i на вариантите y_i . Така получаваме извадка, състояща се от ранговете (x'_i, y'_i) на двойките (x_i, y_i) .

Коефициентът $r_{x'y'}$ на корелация на извадката (x'_i, y'_i) се нарича рангов коефициент r_s на изходната извадка (x_i, y_i) (коефициент на Спирмън).

Пример 34.2. Да се намери ранговият коефициент на корелация на извадката

x_i	12,7	10,2	13,6
y_j	2,7	3,2	3,0

Решение: Образуваме вариационния ред на X : 10,2; 12,7; 13,6. Следователно, поредните номера на вариантите $x_1=12,7$, $x_2=10,2$, $x_3=13,6$ във вариационния ред са съответно 2, 1 и 3, т.е. $x'_1=2$, $x'_2=1$, $x'_3=3$.

По същия начин определяме, че $y'_1=1$, $y'_2=3$, $y'_3=2$. Така получаваме ранговата извадка

x'_i	2	1	3
y'_i	1	3	2

за която пресмятаме коефициента на корелация в следната последователност:

$$1) \bar{x}' = \bar{y}' = \frac{1}{3}(1+2+3) = 2,$$

$$2) s_{x'}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{3}[(2-2)^2 + (1-2)^2 + (3-2)^2] = \frac{2}{3},$$

$$s_{y'}^2 = s_{x'}^2 = \frac{2}{3}; \quad s_{x'} = s_{y'} = \sqrt{\frac{2}{3}};$$

$$3) s_{x'y'} = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{3}[(2-2)(1-2) + (1-2)(3-2) + (3-2)(2-2)] = -\frac{1}{3}.$$

$$\text{Следователно, } r_s = r_{x'y'} = \frac{-1/3}{\sqrt{2/3}\sqrt{2/3}} = -0,5. \blacklozenge$$

Едно от предимствата на ранговия коефициент на корелация е лесното му пресмятане. Доказва се, че не е необходимо да се прилага формула (34.2), тъй като

$$r_s = 1 - \frac{6}{n(n-1)} \sum_{i=1}^n (x'_i - y'_i)^2.$$

По тази формула за ранговия коефициент от пример 2 имаме

$$r_s = 1 - \frac{6}{3(3^2-1)}[(2-1)^2 + (1-3)^2 + (3-2)^2] = 1 - \frac{1}{4} \cdot 6 = -0,5.$$

Доказва се също, че ако между X и Y има линейна зависимост, т.е. $aX + bY + c = 0$, то $|\rho_{xy}| = |r_{xy}| = |r_s| = 1$.

Забележка 34.1. За намиране на числените характеристики при интервално статистическо разпределение на извадката интервалите се заменят със средите им.

Забележка 34.2. Пресмятането на числените характеристики се улеснява с помощта на линейните трансформации:

$$x_i = au_i + b, \quad y_i = cv_i + d.$$

$$\text{Тогава: } \bar{x} = a\bar{u} + b, \quad \bar{y} = c\bar{v} + d, \quad (34.3)$$

$$s_x^2 = a^2 s_u^2, \quad s_y^2 = c^2 s_v^2, \quad s_{xy} = a \cdot c \cdot s_{uv}, \quad r_{xy} = r_{uv} \quad (34.3)$$

Пример 34.3. В резултат на 100 измервания е получена

корелационната таблица:

	(x_{i-1}, x_i)	(0,2)	(2,4)	(4,10)
(y_{j-1}, y_j)				
(20,30)	20	—	—	
(30,50)	—	30	10	
(50,120)	—	20	20	

Да се намерят точковите оценки на двумерната случайна величина (X, Y) .

Решение: Заменяме интервалите със средите им:

x_i	1	3	7	m_{y_j}
y_j				
25	20	0	0	20
40	0	30	10	40
85	0	20	20	40
m_{x_i}	20	50	30	100

$$\text{Полагаме } u_i = x_i - 3, \quad v_j = \frac{1}{5}(y_j - 40) = \frac{1}{5}y_j - 8 \text{ и}$$

получаваме таблицата:

u_i	-2	0	4	m_{v_j}
v_j				
-3	20	0	0	20
0	0	30	10	40
9	0	20	20	40
m_{u_i}	20	50	30	100

, от която пресмятаме:

$$1) \bar{u} = \frac{1}{100}(-2 \cdot 20 + 0 + 4 \cdot 30) = 0,8, \quad \bar{u}^2 = \frac{1}{100}(4 \cdot 20 + 16 \cdot 30) = 5,6,$$

$$s_u^2 = \overline{u^2} - \bar{u}^2 = 4,96, \quad s_u = \sqrt{4,96} = 2,23.$$

$$2) \bar{v} = \frac{1}{100}(-3.20 + 0 + 9.40) = 3, \quad \bar{v}^2 = \frac{1}{100}(9.20 + 81.40) = 34,20,$$

$$s_v^2 = \bar{v}^2 - \bar{v}^2 = 25,20, \quad s_v = \sqrt{25,20} = 5,02.$$

$$3) \overline{xy} = \frac{1}{100}[(-2)(-3).20 + 0 + 0 + 0 + 0 + 0.0.30 + 0 + 0 + 0 + 9.4.20] = 8,4,$$

$$s_{uv} = 8,4 - 0,8.3 = 6;$$

$$4) r_{uv} = \frac{6}{2,23 \cdot 5,02} = 0,536.$$

От формули (34.3), (34.4): $x_i = u_i + 3 \Rightarrow \bar{x} = 0,8 + 3 = 3,8, \quad s_x^2 = s_u^2 = 4,96,$
 $y_i = 5v_i + 40 \Rightarrow \bar{y} = 5.3 + 40 = 55, \quad s_y^2 = 25.s_v^2 = 630,$
 $s_{xy} = 1.5.s_{uv} = 30 \quad r_{xy} = r_{uv} = 0,536.$

Така получаваме следните точкови оценки:

- За математическите очаквания EX и EY $\bar{x} = 3,8, \quad \bar{y} = 55.$
- За дисперсиите DX и DY – поправените дисперсии $\tilde{s}_x^2 = \frac{100}{100-1} s_x^2 = 6,06$ и $\tilde{s}_y^2 = \frac{5}{4} 630 = 787,5.$
- За ковариацията $\text{cov}(X, Y)$ – извадковата ковариация $\tilde{s}_{xy} = \frac{100}{99} s_{xy} = 30,30.$
- За коефициента на корелация ρ_{XY} – извадъчният коефициент на корелация $r_{xy} = 0,536.$ ♦

§35. Елементи на корелационния анализ. Видове зависимости между две величини.

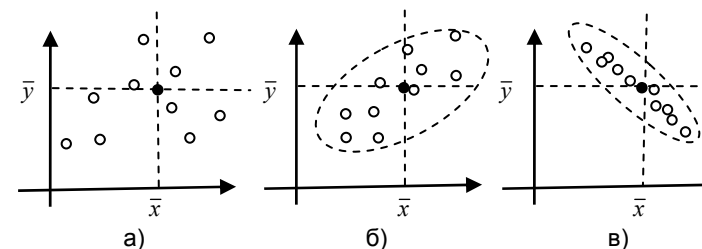
Нека за изучаване на признаците X и Y на генералната съвкупност са направени n наблюдения. Разглеждайки (X, Y) като случайна величина, а получените данни $(x_1, y_1), \dots, (x_n, y_n)$ като нейни реализации, си поставяме задачата да изследваме зависимостта между компонентите X и Y .

1. Видове зависимости. Две случайни величини X и Y се наричат независими, ако законът на едната величина не зависи от това каква стойност е приела другата. В такъв случай диаграмата на разсейването има вида на фиг. 35.1а.

Две величини могат да бъдат свързани с функционална зависимост, например $Y = X^2, \quad Y = aX + b$ - линейна зависимост и т.н.

Всички останали зависимости са статистически зависимости. Предмет на изследване тук ще бъде така наречената корелационна

зависимост (зависимост на тенденциите) (фиг.35.1б,в) – зависимост при която при нарастване на X се забелязва нарастване (положителна корелация) на стойностите на Y или намаляване (отрицателна корелация).



Фиг.35.1

Както вече показавме (виж. §17), за независимите величини $\text{cov}(X, Y) = 0$, но не и обратното. Величини, за които $\text{cov}(X, Y) = 0$, се наричат некорелирани, а за които $\text{cov}(X, Y) \neq 0$ - корелирани.

Основни инструменти за изследване на корелацията на две величини са извадъчната ковариация s_{xy} и особено извадъчният коефициент на корелация r_{xy} , които са оценки на съответно на $\text{cov}(X, Y)$ и ρ_{XY} . Например:

- на фиг.35.1б повече от произведенията $(x_i - \bar{x})(y_i - \bar{y})$ са положителни, откъдето $s_{xy} = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y}) > 0$, а и от там $r_{xy} > 0$ (положителна корелация)
- на фиг.35.1в е представен случая $s_{xy} < 0, \quad r_{xy} < 0$ (отрицателна корелация)
- за фиг.35.1а) s_{xy} и r_{xy} са близки до нула (некорелирани величини).

Корелационният анализ е дял от статистиката, който изследва корелационната зависимост между две величини. Основни задачи:

- Намиране на оценки на коефициента на корелация ρ_{XY} .
- Доказване на наличието или отсъствието на корелационна зависимост.
- Изследване на силата на линейната връзка между величините.

2. Доверителни интервали за коефициента на корелация ρ_{XY} на нормално разпределени случайни величини X и Y .

Нека по дадена двумерна извадка на два нормално разпределени признака X и Y на генералната съвкупност и нека е изчислена оценката

r_{xy} на коефициента на корелация ρ_{XY} . За определяне на доверителния интервал за ρ_{XY} се използва, че разпределението на ρ_{XY} е близко до нормално разпределение $N(m^*, \sigma^*)$, с параметри

$$m^* = \frac{1}{2} \ln \frac{1+r_{xy}}{1-r_{xy}}, \quad \sigma^* = \frac{1}{\sqrt{n-3}}.$$

Намирането на **доверителния интервал** за ρ_{XY} **доверителна вероятност γ** се извършва в следната последователност:

- изчисляваме $m^* = \frac{1}{2} \ln \frac{1+r_{xy}}{1-r_{xy}}$ и квантилът $Z_{\frac{1+\gamma}{2}}$ на $Z \sim N(0,1)$;

- изчисляваме $z_1 = m^* - Z_{\frac{1+\gamma}{2}} \cdot \frac{1}{\sqrt{n-3}}$ и $z_2 = m^* + Z_{\frac{1+\gamma}{2}} \cdot \frac{1}{\sqrt{n-3}}$;

- Определяме доверителния интервал $\rho_{xy} \in \left(\frac{e^{2z_1} - 1}{e^{2z_1} + 1}, \frac{e^{2z_2} - 1}{e^{2z_2} + 1} \right)$

Посоченото правило за намиране на доверителен интервал може да се приложи и за оценка на ранговия коефициент на корелация.

Освен като интервална оценка за ρ_{XY} , с помощта на доверителния интервал могат да се проверяват предположения за вида на зависимостта между X и Y :

- Ако доверителният интервал съдържа числото нула, то с вероятност γ може да твърдим, че величините X и Y не са корелирани;
- Ако доверителният интервал съдържа числата 1 или -1, то с вероятност γ може да се твърди, че между X и Y има силна линейна връзка.

3. Проверка на хипотезата за значимостта на корелационната зависимост между нормално разпределени величини X и Y .

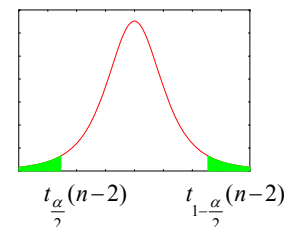
Разглеждаме хипотезата H_0 за некорелираност на величините X и Y , т.е. $H_0 = \{\rho_{XY} = 0\}$.

Статистическият критерий, с който се проверява тази хипотеза, е t -разпределението на Стюдънт с $n-2$ степени на свобода $t(n-2)$, като

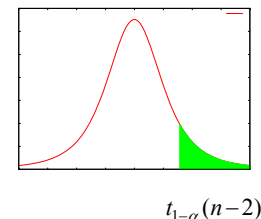
$$\text{наблюдаваната стойност е } t_{\text{набл.}} = \frac{r_{xy} \sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

Критичната област се определя в зависимост от конкуриращата хипотеза и избраното ниво на значимост α :

- Ако $H_1 = \{\rho_{XY} \neq 0\}$:

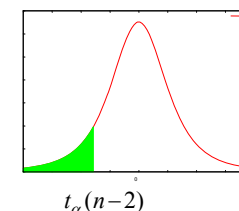


- Ако



$$H_1 = \{\rho_{XY} < 0\}$$

- Ако $H_1 = \{\rho_{XY} > 0\}$



Така, като се използва четността на плътността на t -разпределението, т.е. $t_{\alpha}(n-1) = -t_{1-\alpha}(n-1)$, проверката на хипотезата H_0 при ниво на значимост α по дадена извадка с обем n се извършва по следния начин:

$$1) \text{ Изчислява се } t_{\text{набл.}} = \frac{r_{xy} \sqrt{n-2}}{\sqrt{1-r_{xy}^2}};$$

2) От таблицата за квантилите на t -разпределението с $n-1$ степени на свобода съгласно конкуриращата хипотеза се определя квантилът:

- $t_{\text{кр}} = t_{1-\alpha/2}(n-2)$ при $H_1 = \{\rho_{XY} \neq 0\}$. Област на приемане на H_0 е

$$|t_{\text{набл.}}| < t_{\text{кр}}.$$

- $t_{\text{кр}} = t_{1-\alpha}(n-2)$ при $H_1 = \{\rho_{XY} > 0\}$. Област на приемане на H_0 е $t_{\text{набл.}} < t_{\text{кр}}$

- $t_{\text{кр}} = -t_{1-\alpha}(n-2)$ при $H_1 = \{\rho_{XY} < 0\}$. Област на приемане на H_0 е $t_{\text{набл.}} > t_{\text{кр}}$.

Приемането на хипотезата означава, че с вероятност $1-\alpha$ между няма корелационна зависимост, а отхвърлянето – че има корелационна зависимост.

Същият тест се прилага и при проверка на хипотезата за значимост на ранговия коефициент на корелация, т.е. за отхвърляне или приемане на рангова корелационна зависимост

Пример. 35.1. В конни състезания състезателните коне, които са номерирани съгласно ръста си, са заели следните места: 6, 5, 1, 4, 2, 7, 8, 10, 3, 9. С ниво на значимост $\alpha=0,05$ да се провери хипотезата, че няма рангова корелационна зависимост между ръста и мястото, което състезателният кон е заел ($H_0 = \{r_s = 0\}$).

Решение. Означаваме с x_i номерата на конете съгласно ръста им, а с y_i - местата им в класирането. Очевидно, x_i са ранговете на теглата на конете, а y_i - ранговете на времената им за изминаване на разстоянието.

Нанасяме данните в таблицата

x_i	1	2	3	4	5	6	7	8	9	10
y_i	6	5	1	4	2	7	8	10	3	9

Пресмятаме ранговия коефициент:

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^n (x_i - y_i)^2 = 1 - \frac{6}{990} [5^2 + 3^2 + 2^2 + 3^2 + 1^2 + 1^2 + 2^2 + 6^2 + 1^2] = 0,4545$$

Изчисляваме наблюдаваната стойност:

$$t_{\text{набл.}} = \frac{r_s \sqrt{n-2}}{\sqrt{1-r_s^2}} = \frac{0,45 \cdot \sqrt{8}}{\sqrt{1-0,45^2}} = \frac{1,273}{0,893} = 1,42.$$

От таблицата за t -разпределението изчисляваме

$$t_{\text{кр.}} = t_{1-\frac{\alpha}{2}}(10-2) = t_{0,975}(8) = 2,31.$$

Тъй като $|t_{\text{набл.}}| < t_{\text{кр.}}$, то нямаме основание да отхвърлим хипотезата H_0 , т.е. няма рангова корелационна зависимост между ръста и мястото на класиране.

Упражнения.

1. За изследването на променливите X и Y е получена извадката

x_i	-1	2	2	3	6
y_i	0	-1	2	4	8

Да се представят графически резултатите, да се намери точкова и интервална оценка с доверителна вероятност 0,99 на коефициента на корелация.

2. Направени са следните наблюдения

X	2	4	7	6	8
Y	2	12	16	18	21

на променливите

X и Y , за които се предполага, че са нормално разпределени случайни величини. Да се провери хипотезата за корелираност на величините.

36. Елементи на регресионния анализ. Метод на най-малките квадрати за намиране на линията на регресия.

Регресионният анализ изучава зависимости между две непрекъснати величини X и Y като, за разлика от корелационния анализ, се прилага при изучаване на зависимости, близки до функционалните, т.е. когато едната величина (Y) се влияе от стойността, която е приела другата величина. Величината X може и да не бъде случайна. За величината Y предполагаем, че има случаен характер, обусловен от грешки на измерването или от други неизвестни причини.

Много често зависимостта на Y , се изразява в това, че средната й стойност зависи от стойността x , която е приела величината X , т.е.

$$E(Y|X=x) = E(Y|x) = f(x),$$

където $f(x)$ е функция, дефинирана за всяка възможна стойност x на величината X . В такъв случай казваме, че Y зависи регресионно от X , а функцията $f(x)$ се нарича регресия на Y от X . (виж §18). Тогава може да представим Y във вида

$$Y = f(X) + E, \quad (36.1)$$

където E случайна величина, отразяваща отклонението на Y от средната й стойност.

Случаи на регресионна зависимост са:

- средната стойност на кръвното налягане от възрастта на индивида.
- количеството продадена минерална вода от температурата,
- средния добив от декар от количеството на валежите.

Основни задачи на регресионния анализ:

- Определяне на вида на зависимостта между променливите.
- Намиране на функцията на регресия.
- Определяне на доверителни интервали и проверка на надеждността на регресионния модел.
- Прилагане на регресионния модел за предвиждане на изменението на Y .

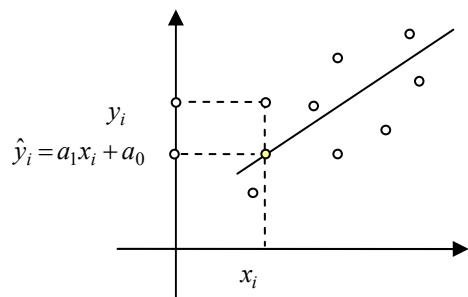
Уравнение на линейната регресия. Да предположим, че регресионната зависимост на Y от X е линейна, т.е. $f(x) = \alpha_1 x + \alpha_0$.

Нека е получена извадката (x_i, y_i) , $(i=1, \dots, n)$. Тогава математическото очакване на Y при $X = x_i$ е $E(Y|x_i) = \alpha_1 x_i + \alpha_0$, а

$$\varepsilon_i = y_i - (\alpha_1 x_i + \alpha_0) \quad (36.2)$$

е отклонението на наблюдаваната стойност y_i от него (фиг.36.1), което съгласно предположението се дължи на случайни фактори. Като приемаме, че е малко вероятно ε_i да бъде голямо, търсим такава линейна функция $y = \alpha_1 x + \alpha_0$, която най-добре да апроксимира данните,

т.е. да минимизира разликите (36.2). Коефициентите a_1 и a_0 ще бъдат оценки на коефициентите α_1 и α_0 на линейната регресия.



Фиг. 36.1.

Намирането на линейната функция $y = a_1x + a_0$ ще извършим по метода на най-малките квадрати, който се състои в следното:

За произволни a_1 и a_0 образуваме сумата от квадратите на разстоянията (36.2) между точката (x_i, y_i) и точката $(x_i, a_1x_i + a_0)$ от правата:

$$S(a_1, a_0) = \sum_{i=1}^n (a_1x_i + a_0 - y_i)^2.$$

Коефициентите a_1 и a_0 определяме като стойностите, при които $S(a_1, a_0)$ има минимум, условието за което е частните производни на функцията $S(a_1, a_0)$ спрямо променливите a_1 и a_0 да се анулират, т.е.

$$\begin{cases} \frac{\partial S}{\partial a_1} = 2 \sum_{i=1}^n (a_1x_i + a_0 - y_i) \cdot x_i = 0 \\ \frac{\partial S}{\partial a_0} = 2 \sum_{i=1}^n (a_1x_i + a_0 - y_i) = 0 \end{cases} \Leftrightarrow \begin{cases} a_1 \sum_{i=1}^n x_i^2 + a_0 \sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i \\ a_1 \sum_{i=1}^n x_i + na_0 = \sum_{i=1}^n y_i \end{cases}$$

Като разделим всяко от уравненията на n , получаваме $\begin{cases} a_1 \overline{x^2} + a_0 \bar{x} = \overline{xy} \\ a_1 \bar{x} + a_0 = \bar{y} \end{cases}$.

От второто уравнение изразяваме $a_0 = \bar{y} - a_1 \bar{x}$, а от първото намираме

$$a_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}.$$

Следователно,

уравнението на линейната регресия на Y от X е

$$y = a_1x + a_0, \quad \text{където } a_1 = \frac{S_{xy}}{S_x^2}, \quad a_0 = \bar{y} - a_1 \bar{x}. \quad (36.3)$$

Уравнението на линейната регресия се преобразува по следния начин:

$$y - \bar{y} = \frac{S_{xy}}{S_x^2} (x - \bar{x}) \Leftrightarrow \frac{y - \bar{y}}{S_y} = r_{xy} \frac{x - \bar{x}}{S_x}, \quad r_{xy} = \frac{S_{xy}}{S_x S_y},$$

където r_{xy} е коефициентът на извадъчната корелация. От тук следва, че

регресионната права минава през центъра $(\bar{x}, \bar{y})$ на извадката, а $a_1 = \frac{S_{xy}}{S_x^2} = \text{tg } \varphi$ е ъгловият ѝ коефициент (характеризира наклона на правата).

По същия начин се получава и уравнението на линейната регресия на X от Y , като тук получаваме уравнението във вида

$$x = b_1y + b_0, \quad b_1 = \frac{S_{xy}}{S_y^2}, \quad b_0 = \bar{x} - b_1 \bar{y},$$

и то може да бъде записано и във вида $\frac{x - \bar{x}}{S_x} = r_{xy} \frac{y - \bar{y}}{S_y}$.

Методът на най-малките квадрати се използва за намиране на линията на регресия и в общия случай, когато се търси друг вид зависимост между X и Y . Видът на зависимостта $y = f(x)$ предварително се определя, например,

$$f(x) = a_0 + a_1x + a_2x^2, \quad f(x) = a_0 + a_1 \sin x + a_2 \cos x, \quad f(x) = a_0 + a_1 e^x, \text{ т.е.}$$

$f(x) = f(x, a_0, a_1, \dots, a_s)$, където неизвестните коефициенти $a_0, a_1, \dots, a_s$ се определят от като точката $(a_0, a_1, \dots, a_s)$, за която функцията

$$S(a_0, a_1, \dots, a_s) = \sum_{i=1}^n [y_i - f(x_i, a_0, \dots, a_s)]^2$$

има минимум. Системата, от която определяме $a_0, a_1, \dots, a_s$, е

$$\frac{\partial S}{\partial a_0} = 0, \quad \frac{\partial S}{\partial a_1} = 0, \dots, \quad \frac{\partial S}{\partial a_s} = 0.$$

Например, за намиране на квадратна регресия $f(x) = a_0 + a_1x + a_2x^2$ на Y от X , коефициентите a_0, a_1, a_2 се определят от системата:

$$\begin{cases} a_0 n + a_1 \sum x_i + a_2 \sum x_i^2 = \sum y_i \\ a_0 \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3 = \sum x_i y_i \\ a_0 \sum x_i^2 + a_1 \sum x_i^3 + a_2 \sum x_i^4 = \sum x_i^2 y_i \end{cases} \quad (36.4)$$

Пример. 36.1. Установено е, че шурците свирят, като бързо търкат крилцата си. Изследва се зависимостта между честотата X (брой на трептения за сек) и температурата Y (в градуси) на въздуха. Получена е извадката:

(20, 31,67), (16, 22,22), (20, 33,89),
 (18, 28,89), (17, 27,22), (16, 23,89),
 (15, 21,11), (17, 27,78), (15, 20,56),
 (16, 28,33), (15, 26,67), (17, 28,33),
 (16, 27,22), (17, 28,89), (14, 24,44).

Да се намери уравнението на линейната регресия. Ако имаме уред за мерене на честотата на трептенията и той отчита $x=19$, колко градуса е температурата на въздуха?

Решение. Изчисленията са представени в таблицата, последният ред на която е получен чрез сумиране по стълбове.

x_i	y_i	x_i^2	$x_i y_i$
20	31,67	400	633,4
16	22,22	256	355,52
20	33,89	400	677,8
18	28,89	324	520,02
17	27,22	289	462,74
16	23,89	256	382,24
15	21,11	225	316,65
17	27,78	289	472,26
15	20,56	225	308,4
16	28,33	256	453,28
15	26,67	225	400,05
17	28,33	289	481,61
16	27,22	256	435,52
17	28,89	289	491,13
14	24,44	196	342,16
249	401,11	4175	6732,78

Тогава:

$$\bar{x} = \frac{249}{15} = 16,6, \quad \bar{y} = \frac{401,11}{15} = 26,74.$$

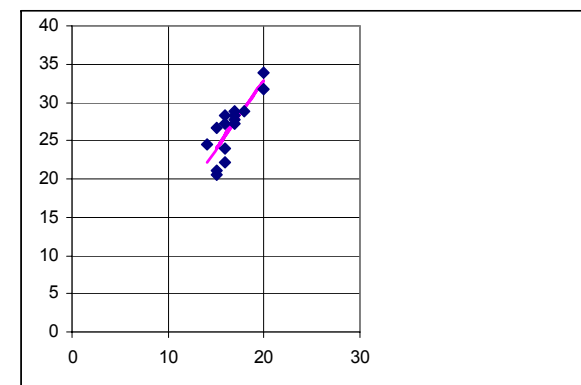
$$s_x^2 = \overline{x^2} - (\bar{x})^2 = \frac{4175}{15} - 16,6^2 = 2,78.$$

$$s_{xy} = \overline{xy} - \bar{x} \cdot \bar{y} = \frac{6732,78}{15} - 16,6 \cdot 26,74 = 13,00$$

Следователно, $a_1 = \frac{13}{2,78} = 4,67$, $a_0 = 26,74 - 4,67 \cdot 16,6 = -50,93$, откъдето

$$y = 4,67x - 50,93. \quad (\text{фиг. 36.2})$$

Като положим $x=19$, получаваме $y=31,03^0$, т.е. може да считаме, че температурата на въздуха е 31^0 .



Фиг. 36.2.

Забележка. 36.1. В пример 36.1. имаме случай, когато и двете изследвани величини са случайни.

Пример 36.2. Да се намери уравнението на квадратичната регресионна зависимост по данните

$$\begin{array}{c|ccccc} x_i & -2 & -1 & 0 & 1 & 2 \\ \hline y_i & 1,2 & 0,1 & -0,8 & 0,2 & 0,8 \end{array}$$

Решение: За получим системата (36.4) за коефициентите a_0, a_1, a_2

на квадратната функция $f(x) = a_0 + a_1 x + a_2 x^2$ извършваме пресмятанията като използваме дадената по-долу таблица:

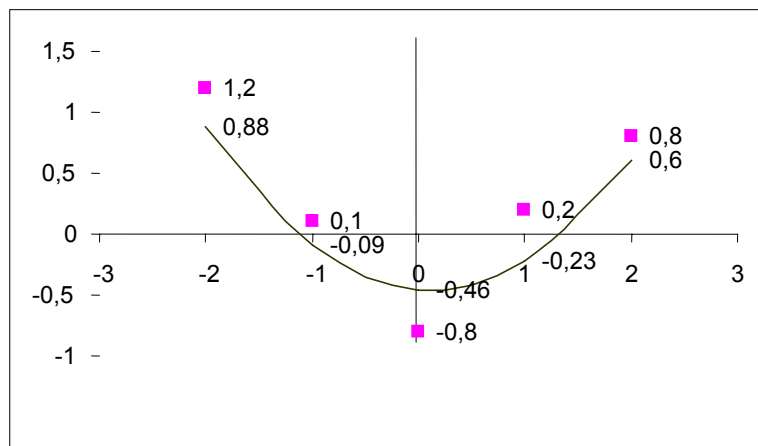
$$\begin{aligned} \sum x_i &= 0, & \sum x_i^2 &= 10, & \sum x_i^3 &= 0, & \sum x_i^4 &= 34 \\ \sum y_i &= 1,5, & \sum x_i y_i &= -0,7, & \sum x_i^2 y_i &= 8,3. \end{aligned}$$

x_i	y_i	x_i^2	x_i^3	x_i^4	$x_i y_i$	$x_i^2 y_i$
-2	1,2	4	-8	16	-2,4	4,8
-1	0,1	1	-1	1	-0,1	0,1
0	-0,8	0	0	0	0	0
1	0,2	1	1	1	0,2	0,2
2	0,8	4	8	16	1,6	3,2
0	1,5	10	0	34	-0,7	8,3

Така получаваме

$$\begin{cases} 5a_0 + 10a_2 = 1,5 \\ 10a_1 = -0,7 \\ 10a_0 + 34a_2 = 8,3 \end{cases} \text{ с решение } a_0 = 0,03, a_1 = -0,07, a_2 = -0,46,$$

Следователно, $y = 0,3x^2 - 0,07x - 0,46$ (фиг.36.3)



фиг.36.3.

37. Статистически анализ на уравнението на линейната регресия.

В този параграф ще разгледаме адекватността (съответствието с наличните данни) на получения регресионен модел.

В най-простия случай на линейна регресионна зависимост на Y от X , т.е. $EY = \alpha_1 X + \alpha_0$ (линейна регресия на Y от X) може да представим Y във вида

$$Y = \alpha_1 X + \alpha_0 + \epsilon, \quad (37.1)$$

(смутено уравнение на права), където ϵ е случайна величина, обусловена от случайни фактори, която определя отклонението на Y от математическото ѝ очакване.

Пример 37.1. Нека Y е физична величина, която зависи линейно от величината X и нека измерваме Y при различни (неслучайни) стойности на X . Тогава:

- ϵ е грешката от измерването. Ако измерването се провежда с един и същ уред, точността на който не зависи от стойността на X и не се правят систематични грешки, то ϵ е случайна величина с нормално разпределение $N(0, \sigma)$ като средното квадратично отклонение не зависи от X и се определя от точността на измервателния уред.
- законът на величината Y е $N(EY, \sigma)$ като, съгласно предположението за съществуване на линейна регресионна зависимост, имаме $EY = \alpha_1 X + \alpha_0$.
- за предвидена стойност на Y при $X = x^*$ приемаме стойността $y^* = \alpha_1 x^* + \alpha_0$. Тази стойност съпада с условното математическо очакване на Y при $X = x^*$. (фиг. 37.1). ♦

Ако коефициентите α_1 и α_0 не са известни, то по дадена извадка (x_i, y_i) , $i = 1, 2, \dots, n$ намираме оценките им a_1 и a_0 съгласно формули (36.3). Очевидно, може да запишем аналогично на (37.1) съотношение (фиг.36.1) $y_i = a_1 x_i + a_0 + \epsilon_i$, или $y_i = \hat{y}_i + \epsilon_i$, където

$$\hat{y}_i = a_1 x_i + a_0 \quad (37.2)$$

е предвидената стойност на Y при $X = x_i$ според регресионния модел.

Ще получим извадъчните параметри на разпределението на величините $\hat{y}_i$ - предвидените (според регресионния модел) стойности и ϵ_i - разликата между предвиденияте и наблюдаваните стойности.

1. Събираемост $\epsilon_i = y_i - \hat{y}_i$ отразява грешката, която се прави при замяна y_i с изчислената съгласно модела стойност $\hat{y}_i$ и е наблюдавана стойност на величина ϵ , която се нарича грешка на регресията (на предвиждането). Тази величина се явява следствие както на случайни фактори, така и на неточността на коефициентите a_1 и a_0 .

Като имаме предвид формули (36.2), получаваме:

$$\begin{aligned} \bar{\epsilon} &= \frac{1}{n} \sum \epsilon_i = \frac{1}{n} \sum (y_i - a_1 x_i - a_0) = \frac{1}{n} \sum y_i - \frac{1}{n} \sum a_1 x_i - \frac{1}{n} \sum a_0 = \\ &= \bar{y} - a_1 \bar{x} - a_0 = \bar{y} - a_1 \bar{x} - (\bar{y} - a_1 \bar{x}) = 0, \end{aligned}$$

Извадъчната дисперсия s_{ϵ}^2 се нарича остатъчна дисперсия.

За нея получаваме (заместваме $a_0 = \bar{y} - a_1 \bar{x}$):

$$\begin{aligned}
s_{\varepsilon}^2 &= \frac{1}{n} \sum (\varepsilon_i - \bar{\varepsilon})^2 = \frac{1}{n} \sum \varepsilon_i^2 = \frac{1}{n} \sum (y_i - a_1 x_i - a_0)^2 = \frac{1}{n} \sum [(y_i - \bar{y}) - a_1 (x_i - \bar{x})]^2 \\
&= \frac{1}{n} \sum (y_i - \bar{y})^2 - 2a_1 \cdot \frac{1}{n} \sum (y_i - \bar{y})(x_i - \bar{x}) + a_1^2 \cdot \frac{1}{n} \sum (x_i - \bar{x})^2 \\
&= s_y^2 - 2a_1 s_{xy} + a_1^2 s_x^2 = s_y^2 - 2 \frac{s_{xy}}{s_x^2} s_{xy} + \left(\frac{s_{xy}}{s_x^2} \right)^2 s_x^2 = s_y^2 - \frac{s_{xy}^2}{s_x^2},
\end{aligned}$$

Като заместим $a_1 = \frac{s_{xy}}{s_x^2}$, получаваме следната *фирмула за пресмятане*

на остатъчната дисперсия:

$$s_{\varepsilon}^2 = s_y^2 - a_1 s_{xy} \quad (37.3)$$

2. Извадъчната средна на предвидените стойности $\hat{y}_i$ намираме по формула (37.2):

$$\hat{y} = \bar{y} - \bar{\varepsilon} = \bar{y} \text{ - средна стойност на предвиждането.}$$

За средно квадратичното на $\hat{y}_i$ (дисперсия на предвидените стойности) имаме:

$$s_{\hat{y}}^2 = \frac{1}{n} \sum (\hat{y}_i - \bar{\hat{y}})^2 = \frac{1}{n} \sum (\hat{y}_i - \bar{y})^2 = \frac{1}{n} \sum (a_1 x_i + \bar{y} - a_1 \bar{x} - \bar{y})^2 = a_1^2 \frac{1}{n} \sum (x_i - \bar{x})^2 = a_1^2 s_x^2$$

Тъй като $a_1 = \frac{s_{xy}}{s_x^2}$, то получаваме.

$$s_{\hat{y}}^2 = a_1^2 s_x^2 = a_1 s_{xy} \quad (37.4)$$

Като имаме предвид формула (37.3), заключаваме, че между общата дисперсия s_y^2 , остатъчната дисперсия s_{ε}^2 и дисперсията $s_{\hat{y}}^2$ връзката:

$$s_y^2 = s_{\varepsilon}^2 + s_{\hat{y}}^2$$

Тази формула има следното тълкувание:

Общата извадъчна дисперсия s_y^2 на зависимата величина Y се разпада на две събираеми:

- $s_{\varepsilon}^2 = s_y^2 - a_1 s_{xy}$ - остатъчна дисперсия (дисперсия на грешките на оценките), която е следствие от случайни фактори и грешки от измерванията или на изменението на неотчетени в регресионния модел фактори.
- $s_{\hat{y}}^2 = a_1 s_{xy}$ - факторна дисперсия (дисперсия на предвидените стойности), която характеризира разсейването на предвидените стойности $\hat{y}_i$ на Y около общата средна $\bar{y}$.

Забележка 37.1. За отношението $\frac{s_{\hat{y}}^2}{s_y^2}$, което се нарича коэффициент

на детерминация имаме

$$\frac{s_{\hat{y}}^2}{s_y^2} = \frac{a_1 s_{xy}}{s_y^2} = \frac{s_{xy}^2}{s_y^2 s_x^2} = r_{xy}^2.$$

В по-нататъчните разглеждания ще считаме, че за всяка стойност x на X законът на ε е $\varepsilon \sim N(0, \sigma)$ (не зависи наблюдаваната стойност на X)

При това предположение, ако се въведат оценките:

$$\tilde{s}_{\varepsilon}^2 = \frac{n}{n-2} s_{\varepsilon}^2 \text{ - поправената остатъчна дисперсия на } D\varepsilon = \sigma^2,$$

$$\tilde{s}_{\varepsilon} = \sqrt{\tilde{s}_{\varepsilon}^2} \text{ - стандартна грешка на оценката (на регресията)}$$

$$s_{a_1} = \frac{\tilde{s}_{\varepsilon}}{s_x \sqrt{n}} \text{ - стандартна грешка на наклона } a_1$$

$$s_{a_0} = s_{a_1} \sqrt{\bar{x}^2} \text{ - стандартна грешка на коефициента } a_0$$

се установява, че статистиките $\frac{a_1 - \alpha_1}{s_{a_1}}$ и $\frac{a_0 - \alpha_0}{s_{a_0}}$ имат t -разпределение с $n-2$ степени на свобода. Следователно,

Доверителните интервали с доверителна вероятност γ за коефициентите α_1 и α_0 са съответно:

$$\begin{aligned}
&\left(a_1 - t_{\frac{\gamma+1}{2}}(n-2)s_{a_1}; a_1 + t_{\frac{\gamma+1}{2}}(n-2)s_{a_1} \right) \text{ и} \\
&\left(a_0 - t_{\frac{\gamma+1}{2}}(n-2)s_{a_0}; a_0 + t_{\frac{\gamma+1}{2}}(n-2)s_{a_0} \right),
\end{aligned}$$

където $t_{\frac{\gamma+1}{2}}(n-2)$ е квантилът от ред $\frac{\gamma+1}{2}$ на t -разпределението с $n-1$ степени на свобода.

Доверителен интервал за $X = x^*$, когато $x^* \in (x_{\min}, x_{\max})$ се получава по следния начин:

Доверителният интервал с доверителна вероятност γ за предвидената стойност y^* :

1) От уравнението на линейната регресия се изчислява оценката на предвиждането $\hat{y}^* = a_0 + a_1 x^*$.

2) Изчислява се статистиката

$$s_{y^*} = s_{\varepsilon} \sqrt{\frac{1}{n} \left(1 + \frac{(x^* - \bar{x})^2}{s_x^2} \right)} \quad (\text{стандартна грешка на предвиждането}).$$

3) Намира се квантилът $t_{\frac{\gamma+1}{2}}(n-2)$.

4) Доверителният интервал е $y^* \in \left(\hat{y}^* - t_{\frac{\gamma+1}{2}} s_{y^*}, \hat{y}^* + t_{\frac{\gamma+1}{2}} s_{y^*} \right)$.

Пример. 37.2. По извадка с обем 17 са получени следните резултати:

$$\bar{x} = 55,77, \quad \bar{y} = 59,55, \quad \bar{x}^2 = 3495,41, \quad \bar{y}^2 = 5021,41, \quad \bar{xy} = 4068,53.$$

Да се намерят оценки за параметрите на линейната регресия и доверителните им интервали с доверителна вероятност 0,95.

Решение. Изчисляваме

$$s_x^2 = 3495,41 - 55,77^2 = 358,12, \quad s_y^2 = 5021,41 - 59,55^2 = 1475,26,$$

$$s_{xy} = 4068,53 - 55,77 \cdot 59,55 = 747,42,$$

$$a_1 = \frac{747,42}{358,12} = 1,94, \quad a_0 = 59,55 - 1,94 \cdot 55,77 = -48,62,$$

Уравнението на линейната регресия е $\hat{y} = 1,96x - 48,62$.

$$\text{Пресмятаме още } s_{\varepsilon}^2 = s_y^2 - a_1 s_{xy} = 25,26 \text{ и } \tilde{s}_{\varepsilon}^2 = \frac{n}{n-2} s_{\varepsilon}^2 = 28,63,$$

$$\tilde{s}_{\varepsilon} = \sqrt{28,63} = 5,36, \quad s_{a_1} = \frac{\tilde{s}_{\varepsilon}}{s_x \sqrt{n}} = \frac{5,36}{\sqrt{385,12} \sqrt{17}} = 0,066$$

За определяне на доверителния интервал намираме квантила $t_{\frac{\gamma+1}{2}}(n-2) = t_{0,975}(15) = 2,13$, следователно,

$$\alpha_1 \in (1,94 - 2,13 \cdot 0,07, 1,94 + 2,13 \cdot 0,07).$$

т.е. с вероятност 0,95 действителната стойност на наклонна α_1 е число в интервала (1,79; 2,09).

Пример 37.3. Фирма възнамерява да изразходва 6000 лв за реклама. За да се проучи как ще се отрази това на количеството на продажбите, са събрани следните данни за вложените в реклама пари X (в хиляди лв) и броя на продажбите (в хиляди)

x_i	1	4	6	10	14
y_i	35	44	40	52	53

- а) Да се намери линейната регресия на Y от X ;
 б) Какъв е броят на продажбите без реклама на продукта?
 в) Какво количество продажби могат да се очакват, ако за реклама се вложат 6000 лв?
 г) да се намери доверителен интервал за предвиждания брой с доверителна вероятност 0,95

Решение. Пресмятаме:

$$\bar{x} = 7, \quad \tilde{s}_x^2 = 20,8, \quad \bar{y} = 44,8, \quad \tilde{s}_y^2 = 47,76, \quad s_{xy} = 29,$$

$$a_1 = \frac{s_{xy}}{s_x^2} = \frac{29}{20,8} = 1,395, \quad a_0 = \bar{y} - a_1 \bar{x} = 35,042, \quad \text{с което намираме}$$

уравнението на линейната регресия $\hat{y} = 1,395x - 35,042$.

б) Ако в уравнението на линейната регресия положим $x=0$, то получаваме, че продажбите без реклама са около 35000 броя.

в) Изчисляваме $y^* = 1,395 \cdot 6 + 35,042 = 43,406$. Следователно, очакваните продажби при вложени 6000 лв са 43400 броя.

г) За да намерим стандартната грешка на предвиждането, пресмятаме

$$1) \text{ остатъчната дисперсия: } s_{\varepsilon}^2 = s_y^2 - a_1 s_{xy} = 47,76 - 1,395 \cdot 29,8 = 6,189,$$

$$2) \text{ поправената остатъчна дисперсия } \tilde{s}_{\varepsilon}^2 = \frac{n}{n-2} s_{\varepsilon}^2 = \frac{5}{3} \cdot 6,189 = 10,315,$$

$$3) \text{ стандартната грешка на регресията } \tilde{s}_{\varepsilon} = \sqrt{10,315} = 3,21,$$

4) стандартната грешка на предвиждането:

$$s_{y^*} = s_{\varepsilon} \sqrt{\frac{1}{n} \left(1 + \frac{(x^* - \bar{x})^2}{s_x^2} \right)} = 3,21 \sqrt{\frac{1}{5} \left(1 + \frac{(6-7)^2}{20,8} \right)} = 3,21 \cdot \sqrt{1,25} = 3,589,$$

5) От таблицата за t -разпределението с $n-2=5-2=3$ степени на свобода определяме квантила от ред $\frac{\gamma+1}{2} = \frac{0,95+1}{2} = 0,975$: $t_{0,975}(3) = 3,18$,

6) представителната грешка с доверителна вероятност 0,95 е $\delta = s_{y^*} \cdot t_{0,975}(3) = 3,589 \cdot 3,18 = 12,37$.

7) Доверителният интервал е

$$y^* \in (43,406 - 12,37; 43,406 + 12,37) = (31,036; 55,706).$$

Следователно, очакваните продажби са между 31000 и 55700.

Упражнения (корелация и регресия):

1. За двумерната извадка $\begin{array}{c|c} X & 8 & 10 & 5 & 8 & 9 \\ \hline Y & 1 & 3 & 1 & 2 & 3 \end{array}$.

а) да се изчисли извадъчният коефициент на корелация; б) да се провери хипотезата за корелационна зависимост на Y и X . в) да се намери уравнението на линейната регресия на Y от X ; г) да се изобразят графически данните и линията на регресия.

2. Извършени са 6 наблюдения на променливите X и Y , които са нанесени в таблицата $\begin{array}{c|c} X & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline Y & 1 & 4 & 5 & 5 & 6 & 8 \end{array}$.

а) да се пресметне коефициентът на корелация; б) С ниво на значимост $\alpha=0,05$ се провери хипотезата за значимост на корелационната зависимост между X и Y ($H_0=\{\rho_{xy}=0\}$ при $H_1=\{\rho_{xy}\neq 0\}$).

3. Направени са следните наблюдения на величините X и Y и е получена извадката

X	2	4	5	6	8
Y	2	1	2	5	10

Да се намери уравнението на регресия на Y от X и представи графично. Според получения регресионен модел да се изчисли каква стойност ще приеме Y , ако $X=4,5$.

4. Изследва се зависимостта между количеството тор X , употребено за единица площ и добива Y . Получени са следните данни:

X	0	1	2	3	4	5
Y	15	22	31	40	48	54

Като се използва, че $\sum x_i^2=55$, $\sum y_i^2=8490$, $\sum x_i y_i=666$, да се намери: а) уравнението на линейната регресия на Y от X . б) да се представят графически данните и линията на регресия; г) да се намери коефициентът на корелация и провери хипотезата за корелираност на X и Y с ниво на значимост $\alpha=0,1$.

5. Направени са следните наблюдения на нормално разпределените величини X и Y : $\begin{array}{c|c} X & 2 & 4 & 7 & 6 & 8 \\ \hline Y & 2 & 12 & 16 & 18 & 21 \end{array}$. Да се намери:

а) уравнението на регресия на Y от X ; б) каква стойност ще получи величината Y , ако $X=5$; в) доверителен интервал за получената стойност с доверителна вероятност 0,95.

6. Изследва се зависимостта от броя X на работниците във фирма и стойността (в хиляди левове) на сключените договори. Получена е следната извадка:

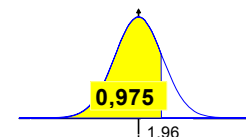
X	1	2	3	4	5	6	7	8	9	10
Y	3	4	3	5	6	8	10	12	11	13

а) да се намери уравнението на линейната регресия на X от Y ; б) да се изчисли извадъчният коефициент на корелация; в) да се изобразят графически данните и линията на регресия.

ПРИЛОЖЕНИЯ

Стойности на функцията $F(x)$ на разпределение на стандартна нормална величина (Z -разпределение)

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt, \quad F(-x) = 1 - F(x)$$



Пример: $F(1,96)=0,975002$

	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,500000	0,503989	0,507978	0,511966	0,515953	0,519939	0,523922	0,527903	0,531881	0,535856
0,1	0,539828	0,543795	0,547758	0,551717	0,555670	0,559618	0,563559	0,567495	0,571424	0,575345
0,2	0,579260	0,583166	0,587064	0,590954	0,594835	0,598706	0,602568	0,606420	0,610261	0,614092
0,3	0,617911	0,621720	0,625516	0,629300	0,633072	0,636831	0,640576	0,644309	0,648027	0,651732
0,4	0,655422	0,659097	0,662757	0,666402	0,670031	0,673645	0,677242	0,680822	0,684386	0,687933
0,5	0,691462	0,694974	0,698468	0,701944	0,705401	0,708840	0,712260	0,715661	0,719043	0,722405
0,6	0,725747	0,729069	0,732371	0,735653	0,738914	0,742154	0,745373	0,748571	0,751748	0,754903
0,7	0,758036	0,761148	0,764238	0,767305	0,770350	0,773373	0,776373	0,779350	0,782305	0,785236
0,8	0,788145	0,791030	0,793892	0,796731	0,799546	0,802337	0,805105	0,807850	0,810570	0,813267
0,9	0,815940	0,818589	0,821214	0,823814	0,826391	0,828944	0,831472	0,833977	0,836457	0,838913
1	0,841345	0,843752	0,846136	0,848495	0,850830	0,853141	0,855428	0,857690	0,859929	0,862143
1,1	0,864334	0,866500	0,868643	0,870762	0,872857	0,874928	0,876976	0,879000	0,881000	0,882977
1,2	0,884930	0,886861	0,888768	0,890651	0,892512	0,894350	0,896165	0,897958	0,899727	0,901475
1,3	0,903200	0,904902	0,906582	0,908241	0,909877	0,911492	0,913085	0,914657	0,916207	0,917736
1,4	0,919243	0,920730	0,922196	0,923641	0,925066	0,926471	0,927855	0,929219	0,930563	0,931888
1,5	0,933193	0,934478	0,935745	0,936992	0,938220	0,939429	0,940620	0,941792	0,942947	0,944083
1,6	0,945201	0,946301	0,947384	0,948449	0,949497	0,950529	0,951543	0,952540	0,953521	0,954486
1,7	0,955435	0,956367	0,957284	0,958185	0,959070	0,959941	0,960796	0,961636	0,962462	0,963273
1,8	0,964070	0,964852	0,965620	0,966375	0,967116	0,967843	0,968557	0,969258	0,969946	0,970621
1,9	0,971283	0,971933	0,972571	0,973197	0,973810	0,974412	0,975002	0,975581	0,976148	0,976705
2	0,977250	0,977784	0,978308	0,978822	0,979325	0,979818	0,980300	0,980774	0,981237	0,981691
2,1	0,982136	0,982571	0,982997	0,983414	0,983823	0,984222	0,984614	0,984997	0,985371	0,985738
2,2	0,986097	0,986447	0,986791	0,987126	0,987455	0,987776	0,988089	0,988396	0,988696	0,988989
2,3	0,989276	0,989556	0,989830	0,990097	0,990358	0,990613	0,990869	0,991106	0,991344	0,991576
2,4	0,991802	0,992024	0,992240	0,992451	0,992656	0,992857	0,993053	0,993244	0,993431	0,993613
2,5	0,993790	0,993963	0,994132	0,994297	0,994457	0,994614	0,994766	0,994915	0,995060	0,995201
2,6	0,995339	0,995473	0,995604	0,995731	0,995855	0,995975	0,996093	0,996207	0,996319	0,996427
2,7	0,996533	0,996636	0,996736	0,996833	0,996928	0,997020	0,997110	0,997197	0,997282	0,997365
2,8	0,997445	0,997523	0,997599	0,997673	0,997744	0,997814	0,997882	0,997948	0,998012	0,998074
2,9	0,998134	0,998193	0,998250	0,998305	0,998359	0,998411	0,998460	0,998511	0,998559	0,998605
3	0,998650	0,998694	0,998736	0,998777	0,998817	0,998856	0,998893	0,998930	0,998965	0,998999
3,1	0,999032	0,999065	0,999096	0,999126	0,999155	0,999184	0,999211	0,999238	0,999264	0,999289
3,2	0,999313	0,999336	0,999359	0,999381	0,999402	0,999423	0,999443	0,999462	0,999481	0,999499
3,3	0,999517	0,999534	0,999550	0,999566	0,999581	0,999596	0,999610	0,999624	0,999638	0,999651
3,4	0,999663	0,999675	0,999687	0,999698	0,999709	0,999720	0,999730	0,999740	0,999749	0,999758
3,5	0,999767	0,999776	0,999784	0,999792	0,999800	0,999807	0,999815	0,999822	0,999828	0,999835
3,6	0,999841	0,999847	0,999853	0,999858	0,999864	0,999869	0,999874	0,999879	0,999883	0,999888
3,7	0,999892	0,999896	0,999900	0,999904	0,999908	0,999912	0,999915	0,999918	0,999922	0,999925
3,8	0,999928	0,999931	0,999933	0,999936	0,999938	0,999941	0,999943	0,999946	0,999948	0,999950
3,9	0,999952	0,999954	0,999956	0,999958	0,999959	0,999961	0,999963	0,999964	0,999966	0,999967

Квантили на разпределението на Стюдънт (t-разпределение)

Пример: $t_{0,975}(8)=2,306004133$ - квантил от ред $p=0,975$ на t-разпределение с $n=8$ степени на свобода

p n	0,900	0,950	0,975	0,990	0,995
1	3,077683537	6,313751514	12,70620473	31,82051595	63,65674115
2	1,885618083	2,91998558	4,30265273	6,964556734	9,9248432
3	1,637744352	2,353363435	3,182446305	4,540702858	5,840909309
4	1,533206273	2,131846782	2,776445105	3,746947388	4,604094871
5	1,475884037	2,015048372	2,570581835	3,364929997	4,032142983
6	1,439755747	1,943180274	2,446911846	3,142668403	3,70742802
7	1,414923928	1,894578604	2,364624251	2,997951566	3,499483297
8	1,39681531	1,859548033	2,306004133	2,896459446	3,355387331
9	1,383028739	1,833112923	2,262157158	2,821437921	3,249835541
10	1,372183641	1,812461102	2,228138842	2,763769458	3,169272672
11	1,363430318	1,795884814	2,200985159	2,718079183	3,105806514
12	1,356217334	1,782287548	2,178812827	2,680997992	3,054539586
13	1,350171289	1,770933383	2,160368652	2,650308836	3,012275833
14	1,345030375	1,761310115	2,144786681	2,624494064	2,976842734
15	1,340605608	1,753050325	2,131449536	2,60248029	2,946712883
16	1,336757167	1,745883669	2,119905285	2,583487179	2,920781621
17	1,33337939	1,739606716	2,109815559	2,566933975	2,898230518
18	1,330390944	1,734063592	2,100922037	2,552379618	2,878440471
19	1,327728209	1,729132792	2,09302405	2,539483189	2,860934604
20	1,325340707	1,724718218	2,085963441	2,527977001	2,845339707
21	1,323187874	1,720742871	2,079613837	2,517648014	2,831359554
22	1,321236742	1,717144335	2,073873058	2,50832455	2,818756056
23	1,31946024	1,713871517	2,068657599	2,499866736	2,807335678
24	1,317835934	1,710882067	2,063898547	2,492159469	2,796939498
25	1,316345073	1,708140745	2,059538536	2,48510717	2,787435805
26	1,314971864	1,705617901	2,055529418	2,478629817	2,778714523
27	1,313702913	1,703288423	2,051830493	2,472659904	2,770682946
28	1,312526782	1,701130908	2,048407115	2,467140089	2,763262442
29	1,311433647	1,699126996	2,045229611	2,46202135	2,756385902
30	1,310415025	1,697260851	2,042272449	2,457261531	2,749995652
40	1,303077053	1,683851014	2,02107537	2,423256774	2,704459262
50	1,298713694	1,675905026	2,008559072	2,403271907	2,677793261
60	1,295821093	1,670648865	2,000297804	2,390119457	2,660283014
70	1,293762898	1,66691448	1,994437086	2,38080746	2,647904603
80	1,292223583	1,664124579	1,990063387	2,373868245	2,638690591
100	1,290074761	1,660234327	1,983971466	2,364217356	2,625890514
120	1,288646234	1,6576509	1,979930381	2,357824599	2,617421135

Квантили на χ^2 -разпределение

Пример: $\chi^2_{0,95}(11)=19,67513757$ - квантил от ред $p=0,95$ на χ^2 -разпределение с $n=11$ степени на свобода.

p n	0,025	0,050	0,900	0,950	0,975	0,990	0,999
1	0,000982069	0,454936425	2,705543971	3,841459149	5,02388647	6,634896712	10,82756622
2	0,050635616	1,386294376	4,605170186	5,991464547	7,377758908	9,210340372	13,81551056
3	0,215795283	2,365973893	6,251388457	7,814727764	9,348403568	11,34486668	16,26623615
4	0,484418557	3,356694001	7,77944034	9,487729037	11,14328678	13,27670414	18,46682695
5	0,831211616	4,351460222	9,236356938	11,07049775	12,83250201	15,08627247	20,51500566
6	1,237344247	5,348120843	10,64464068	12,59158724	14,44937534	16,81189383	22,45774449
7	1,689869192	6,345811373	12,01703656	14,06714043	16,01276427	18,47530691	24,32188634
8	2,179730752	7,344121629	13,36156614	15,50731306	17,53454614	20,09023503	26,12448156
9	2,700389522	8,342832783	14,68365662	16,91897762	19,0227678	21,66599433	27,87716488
10	3,246972789	9,34181805	15,98717917	18,30703805	20,48317735	23,20925116	29,58829845
11	3,81574828	10,34099825	17,27500852	19,67513757	21,92004926	24,72497031	31,26413362
12	4,403788517	11,34032282	18,54934779	21,02606982	23,33666416	26,21696731	32,90949041
13	5,008750539	12,33975614	19,81192931	22,3620325	24,73560489	27,68824961	34,52817898
14	5,628726168	13,33927471	21,06414421	23,68479131	26,11894805	29,14123774	36,12327368
15	6,262137817	14,33885982	22,30712958	24,99579013	27,48839286	30,57791417	37,69729823
16	6,907664402	15,33849951	23,54182892	26,29622761	28,84535072	31,99992691	39,25235481
17	7,564186467	16,33818271	24,76903535	27,58711164	30,1910091	33,4086636	40,79021671
18	8,230746229	17,337903	25,9894231	28,86929943	31,52637844	34,80530572	42,31239633
19	8,906516548	18,33765323	27,20357106	30,14352721	32,85232686	36,19086911	43,82019596
20	9,590777511	19,33742983	28,41198058	31,41043286	34,16960691	37,56623475	45,31474662
21	10,28289782	20,33722858	29,61508943	32,67057337	35,47887591	38,93217269	46,79703804
22	10,98232081	21,33704534	30,81328234	33,92443852	36,7807121	40,28936044	48,26794229
23	11,68855204	22,3368793	32,00689967	35,17246163	38,07562727	41,63839812	49,72823246
24	12,40115026	23,33672677	33,19624426	36,4150285	39,36407706	42,97982015	51,17859777
25	13,1197201	24,33658743	34,38158698	37,65248413	40,64646916	44,31410491	52,61965576
26	13,8439051	25,33645925	35,56317121	38,88513865	41,92317015	45,64168268	54,05196237
27	14,57338291	26,3363399	36,74121675	40,11327205	43,19451096	46,96294214	55,47602018
28	15,30786062	27,3362301	37,91592255	41,33713813	44,46079183	48,27823579	56,89228537
29	16,04707179	28,3361282	39,08746978	42,55696777	45,7222858	49,5878845	58,30117346
30	16,79077241	29,33603221	40,25602376	43,77297178	46,97924223	50,89218135	59,70306426
40	24,43303947	39,33534592	51,80505719	55,75847932	59,34170704	63,69073973	73,40195753
50	32,35736408	49,33493921	63,16712082	67,50480652	71,42019524	76,15389117	86,66081523
60	40,48174843	59,33466815	74,39700583	79,08194439	83,29767502	88,37941893	99,60723316
70	48,75756563	69,33447649	85,52704303	90,53122518	95,02318414	100,4251843	112,3169318
80	57,1531735	79,33433312	96,57820347	101,8794741	106,6285676	112,3287926	124,839224
90	65,64661856	89,33422171	107,5650082	113,1452703	118,1358924	124,1163189	137,2083541
100	74,22192813	99,33413538	118,4980039	124,3421137	129,5611969	135,8067231	149,4492527

Критични точки на F-разпределението при ниво на значимост $\alpha=0,01$

Пример: $F_{0,01}(9, 5)=6,057$ - критична точка от ред $\alpha=0,01$ със степени на свобода: $n_1=5$ на числителя и $n_2=9$ на знаменателя.

n_1 n_2	1	2	3	4	5	6	7	8	9	10	12	24
5	16,258	13,27	12,06	11,392	10,97	10,672	10,456	10,289	10,16	10,05	9,888	9,467
6	13,745	10,92	9,78	9,1483	8,746	8,4661	8,26	8,1017	7,976	7,874	7,718	7,313
7	12,246	9,547	8,451	7,8466	7,46	7,1914	6,9928	6,84	6,719	6,62	6,469	6,074
8	11,259	8,649	7,591	7,0061	6,632	6,3707	6,1776	6,0289	5,911	5,814	5,667	5,279
9	10,561	8,022	6,992	6,4221	6,057	5,8018	5,6129	5,4671	5,351	5,257	5,111	4,729
10	10,044	7,559	6,552	5,9943	5,636	5,3858	5,2001	5,0567	4,942	4,849	4,706	4,327
11	9,646	7,206	6,217	5,6683	5,316	5,0692	4,8861	4,7445	4,632	4,539	4,397	4,021
12	9,3302	6,927	5,953	5,412	5,064	4,8206	4,6395	4,4994	4,388	4,296	4,155	3,781
13	9,0738	6,701	5,739	5,2053	4,862	4,6204	4,441	4,3021	4,191	4,1	3,96	3,587
14	8,8616	6,515	5,564	5,0354	4,695	4,4558	4,2779	4,1399	4,03	3,939	3,8	3,427
15	8,6831	6,359	5,417	4,8932	4,556	4,3183	4,1415	4,0045	3,895	3,805	3,666	3,294
16	8,531	6,226	5,292	4,7726	4,437	4,2016	4,0259	3,8896	3,78	3,691	3,553	3,181
17	8,3997	6,112	5,185	4,669	4,336	4,1015	3,9267	3,791	3,682	3,593	3,455	3,084
18	8,289	6,371	6,632	6,6318	7,006	7,0061	7,0061	7,006	7,591	7,591	7,591	
19	8,1849	5,926	5,01	4,5003	4,171	3,9386	3,7653	3,6305	3,523	3,434	3,297	2,925
20	8,096	5,849	4,938	4,4307	4,103	3,8714	3,6987	3,5644	3,457	3,368	3,231	2,859
21	8,0166	5,78	4,874	4,3688	4,042	3,8117	3,6396	3,5056	3,398	3,31	3,173	2,801
22	6,3707	6,371	6,371	6,3707	6,178	6,1776	6,1776	6,1776	6,178	6,178	6,178	6,178
23	7,8811	5,664	4,765	4,2636	3,939	3,7102	3,539	3,4057	3,299	3,211	3,074	2,702
24	7,8229	5,614	4,718	4,2184	3,895	3,6667	3,4959	3,3629	3,256	3,168	3,032	2,659
25	7,7698	5,568	4,676	4,1774	3,855	3,6272	3,4568	3,3239	3,217	3,129	2,993	2,62
26	7,7213	5,526	4,637	4,14	3,818	3,5911	3,421	3,2884	3,182	3,094	2,958	2,585
27	7,6767	5,488	4,601	4,1056	3,785	3,558	3,3882	3,2558	3,149	3,062	2,926	2,552
28	7,6356	5,453	4,568	4,074	3,754	3,5276	3,3581	3,2259	3,12	3,032	2,896	2,522
29	7,5977	5,42	4,538	4,0449	3,725	3,4995	3,3303	3,1982	3,092	3,005	2,869	2,495
30	6,1776	6,632	7,006	7,0061	7,591	7,591	7,591	7,591	7,591	7,591	8,649	8,649
40	7,3141	5,179	4,313	3,8283	3,514	3,291	3,1238	2,993	2,888	2,801	2,665	2,288
60	7,0771	4,977	4,126	3,649	3,339	3,1187	2,953	2,8233	2,719	2,632	2,496	2,115
80	6,9627	4,881	4,036	3,5631	3,255	3,0361	2,8713	2,742	2,637	2,551	2,415	2,032
100	6,8953	4,824	3,984	3,5127	3,206	2,9877	2,8233	2,6943	2,59	2,503	2,368	1,983
120	6,8509	4,787	3,949	3,4795	3,174	2,9559	2,7918	2,6629	2,559	2,472	2,336	1,95
∞	6,6351	4,605	3,782	3,3194	3,017	2,8022	2,6395	2,5115	2,408	2,321	2,185	1,791

Критични точки на F-разпределението при ниво на значимост $\alpha=0,05$

Пример: $F_{0,05}(9, 5)=3,4817$ - критична точка от ред $\alpha=0,05$ със степени на свобода: $n_1=5$ на числителя и $n_2=9$ на знаменателя.

n_1 n_2	1	2	3	4	5	6	7	8	9	10	12	24
5	6,608	5,786	5,41	5,192	5,05	4,95	4,876	4,818	4,773	4,735	4,678	4,527
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099	4,06	4	3,842
7	5,591	4,737	4,347	4,12	3,972	3,866	3,787	3,726	3,677	3,637	3,575	3,411
8	5,318	8,649	7,591	7,006	6,632	6,371	6,178	6,029	5,911	5,814	5,667	5,279
9	5,117	4,257	3,863	3,633	3,482	3,374	3,293	3,23	3,179	3,137	3,073	2,901
10	4,965	4,103	3,708	3,478	3,326	3,217	3,136	3,072	3,02	2,978	2,913	2,737
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,896	2,854	2,788	2,609
12	4,747	3,885	3,49	3,259	3,106	2,996	2,913	2,849	2,796	2,753	2,687	2,506
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832	2,767	2,714	2,671	2,604	2,42
14	4,6	3,739	3,344	3,112	2,958	2,848	2,764	2,699	2,646	2,602	2,534	2,349
15	4,543	3,682	3,287	3,056	2,901	2,791	2,707	2,641	2,588	2,544	2,475	2,288
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657	2,591	2,538	2,494	2,425	2,235
17	4,451	3,592	3,197	2,965	2,81	2,699	2,614	2,548	2,494	2,45	2,381	2,19
18	3,838	4,066	4,066	4,066	4,459	4,459	4,459	4,459	4,459	4,459	4,459	4,459
19	4,381	3,522	3,127	2,895	2,74	2,628	2,544	2,477	2,423	2,378	2,308	2,114
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393	2,348	2,278	2,083
21	4,325	3,467	3,073	2,84	2,685	2,573	2,488	2,421	2,366	2,321	2,25	2,054
22	4,066	3,838	3,838	3,838	3,838	3,838	3,838	3,838	3,838	3,838	3,838	3,838
23	4,279	3,422	3,028	2,796	2,64	2,528	2,442	2,375	2,32	2,275	2,204	2,005
24	4,26	3,403	3,009	2,776	2,621	2,508	2,423	2,355	2,3	2,255	2,183	1,984
25	4,242	3,385	2,991	2,759	2,603	2,49	2,405	2,337	2,282	2,237	2,165	1,964
26	4,225	3,369	2,975	2,743	2,587	2,474	2,388	2,321	2,266	2,22	2,148	1,946
27	4,21	3,354	2,96	2,728	2,572	2,459	2,373	2,305	2,25	2,204	2,132	1,93
28	4,196	3,34	2,947	2,714	2,558	2,445	2,359	2,291	2,236	2,19	2,118	1,915
29	4,183	3,328	2,934	2,701	2,545	2,432	2,346	2,278	2,223	2,177	2,105	1,901
30	3,838	4,066	4,459	4,459	4,459	4,459	4,459	4,459	4,459	4,459	4,459	5,318
40	4,085	3,232	2,839	2,606	2,45	2,336	2,249	2,18	2,124	2,077	2,004	1,793
60	4,001	3,15	2,758	2,525	2,368	2,254	2,167	2,097	2,04	1,993	1,917	1,7
80	3,96	3,111	2,719	2,486	2,329	2,214	2,126	2,056	1,999	1,951	1,875	1,654
100	3,936	3,087	2,696	2,463	2,305	2,191	2,103	2,032	1,975	1,927	1,85	1,627
120	3,92	3,072	2,68	2,447	2,29	2,175	2,087	2,016	1,959	1,911	1,834	1,608
∞	3,842	2,996	2,605	2,372	2,214	2,099	2,01	1,938	1,88	1,831	1,752	1,517