I. Valószínűségelméleti és matematikai statisztikai alapok

1. A szükséges valószínűségelméleti és matematikai statisztikai alapismeretek összefoglalása

Az alkalmazott statisztikai módszerek tárgyalása, amely e kötet célja, feltételezi a valószínűségszámítás és matematikai statisztika alapvető fogalmainak és módszereinek ismeretét. A témakörnek magyar nyelven is tekintélyes és jól használható szakirodalma van (Vincze I.: Matematikai statisztika ipari alkalmazásokkal, 1975; Prékopa A.: Valószínűségelmélet, 1980; Lukács O.: Matematikai statisztika példatár, 1987; Reimann J., 1992; Rényi A., 1966; Meszéna Gy., Ziermann M.: Valószínűségelmélet és matematikai statisztika, 1981; Kröpfl, B. és mts.: Alkalmazott statisztika, 2000). Ebben és a következő fejezetben ezért csak áttekintjük a szükséges alapokat.

Az 1. fejezetben az alapfogalmakról és a gyakran használatos eloszlásokról lesz szó, a 2. fejezet tárgya a statisztikai következtetés, vagyis a hipotézisvizsgálat és a paraméterbecslés.

1.1. Alapfogalmak

Véletlen jelenség

Ha egy gépről lekerülő termékpéldányok valamely jellemzőjét (pl. a konzervdobozokba töltött paradicsomsűrítmény tömegét) megvizsgáljuk, azt tapasztaljuk, hogy a jellemző értékei különbözőek, és ez az ingadozás elkerülhetetlen. Ugyanígy ingadoznak az egy alkatrész (egy példány) valamely geometriai méretére kapott mérési adatok.

Minden jelenséget az okok egy bizonyos rendszere hoz létre. Ha az okok mindegyikét figyelembe tudnánk venni, a jelenség lefolyása azokból egyértelműen levezethető, kiszámítható volna. Ez azonban gyakorlatilag lehetetlen, vagy célszerűtlen, ezért az esetek túlnyomó többségében az ingadozást véletlenszerűnek nevezzük.

Sokaság és minta

Az egy gépről lekerülő alkatrészek méretadatai, a paradicsomkonzervek tömeg-adatai stb. sokaságot alkotnak. A vizsgálatok célja e sokaság megismerése. Mivel az alapsokaság teljes körű vizsgálatát nem lehet, vagy nem lenne gazdaságos elvégezni, ezért vizsgálatainkat csak az összesség egy kiragadott részére, az ún. mintára korlátozzuk. A minta adatai alapján a matematikai statisztika segítségével következtetünk az alapsokaságra.

Véges sokaság elemeinek meghatározása elvileg lehetséges, de esetleg igen nagy munka. A matematikai statisztika alkalmazása ezt szükségtelenné teszi. Végtelen sokaság esetén az egész sokaság elvileg sem mérhető meg. Például gondoljunk egy adott tárgy tömegének meghatározására. A tömegmérés eredménye a tárgy valódi tömegétől a véletlen hibával különbözik. A lehetséges mérési eredmények végtelen sokaságot alkotnak. Ha a tárgyat mérlegre tesszük, s megmérjük a tömegét, ezzel ki-

választottuk a sokaság egy elemét. A mérést többször megismételve véges számú adatot, a mintát kapjuk.

Valószínűségi változó

Azokat a mennyiségeket, amelyeknek értéke nem állandó, hanem esetről esetre más és más lehet, azonban meghatározható, hogy mekkora valószínűséggel esnek megadott határok közé, valószínűségi változóknak nevezzük.

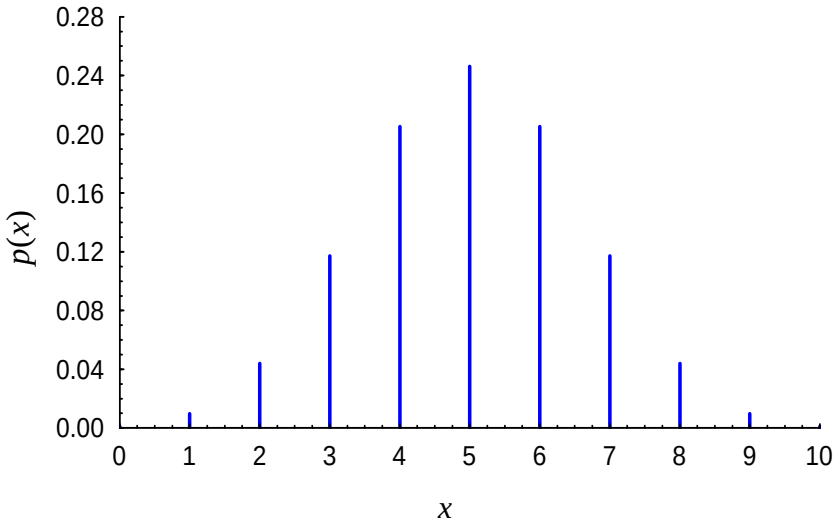
Diszkrét a valószínűségi változó és annak eloszlása, ha egy véges vagy megszámlálhatóan végtelen elemű készletből vehet fel értékeket. Diszkrét valószínűségi változó például az egy műszak alatt gyártott selejtes termékek száma. Lehetséges értékei (0, 1, 2, ..., N) véges sorozatot alkotnak, ahol N az egy műszak alatt gyártott termékek száma. Valamely gyártó gépsor egy műszak alatti üzemzavarainak száma szintén diszkrét valószínűségi változó. Az üzemzavarok lehetséges száma elvileg nem korlátozott, s ha a nagyon nagy számokhoz gyakorlatilag elhanyagolható (igen kicsi) valószínűségeket rendelünk, az üzemzavarok lehetséges száma végtelen sorozatot alkot.

Ha a valószínűségi változó a valós számok folytonos sokaságának értékeit veheti fel, folytonos valószínűségi változóról beszélünk. Folytonos valószínűségi változó pl. az acéltermék szakítószilárdsága, vagy a polimer sűrűsége.

A diszkrét valószínűségi változó sűrűség- és eloszlásfüggvénye

Képzeljük el, hogy egy pénzérmét 10-szer földobunk. Az 1-1*a*) ábrán látható $p(x)$ sűrűségfüggvény "tűi" az egyes $x = k$ értékeknél annak valószínűségét mutatják, hogy a 10 földobás eredménye éppen k-szor fej:

$$p(k) = P(x = k)\,. \qquad (1.1)$$



1-1*a*) ábra. Diszkrét valószínűségi változó sűrűségfüggvénye

A $p(x)$ sűrűségfüggvény tulajdonságai:

$p(x_i) \geq 0$ minden x_i helyen;

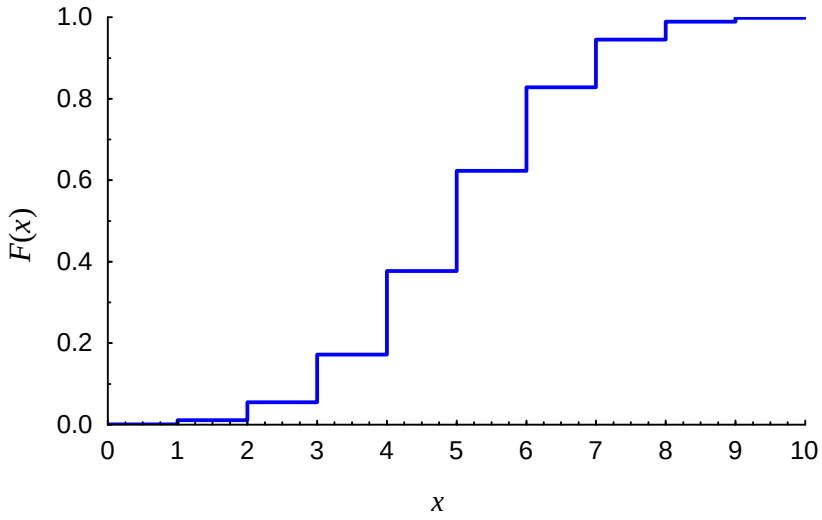
$$\sum_i p(x_i) = 1. \qquad (1.2)$$

A szummázás az összes x_i elemre végzendő.

Szokás a kumulált valószínűségeket is ábrázolni, ezt eloszlásfüggvénynek nevezik. Az 1-1*b*) ábra szerinti $F(x)$ eloszlásfüggvény értéke az $x = k$ helyen azt mutatja, hogy a fej eredményű dobások száma milyen valószínűséggel lesz 10 dobásból legföljebb k:

$$F(k) = P(x \leq k) = \sum_{x_i \leq k} p(x_i). \qquad (1.3)$$

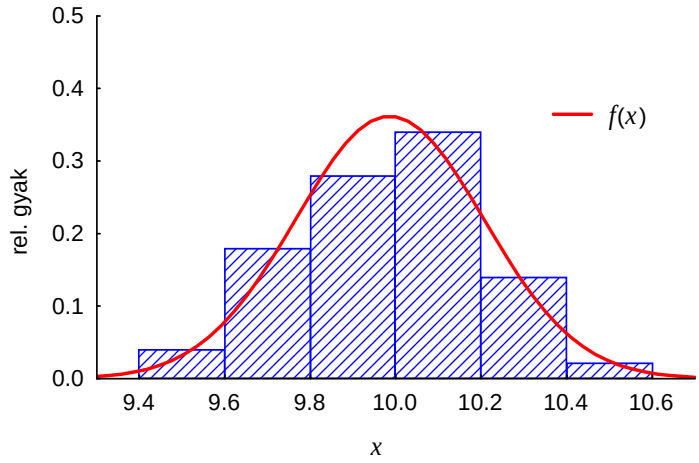
Az irodalomban az $F(k) = P(x < k) = \sum_{x_i < k} p(x_i)$ konvenció is előfordul.



1-1*b*) ábra. Diszkrét valószínűségi változó eloszlásfüggvénye

A folytonos valószínűségi változó sűrűség- és eloszlásfüggvénye

Ábrázoljuk a konkrét mintavétel során kapott értékeket olyan derékszögű koordinátarendszerben, amelynek abszcisszáján a valószínűségi változót osztályokba soroltuk.



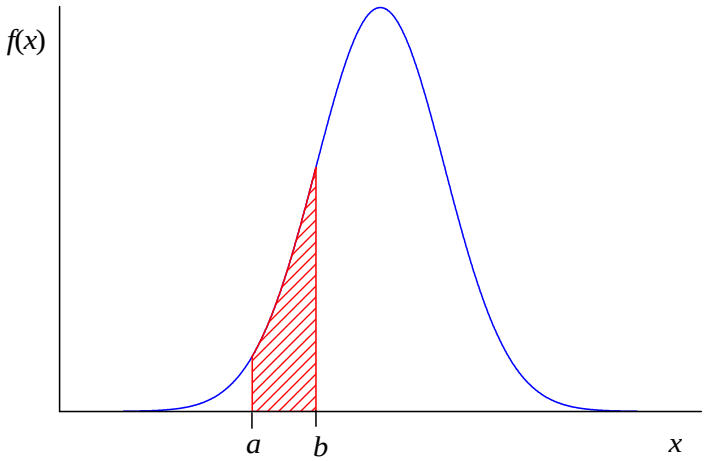
1-2. ábra. Hisztogram és sűrűségfüggvény

A Δx intervallum az osztály szélessége, x_i pedig az osztály közepe, az ún. osztályindex. Az intervallumok mindegyike fölé téglalapot rajzolunk úgy, hogy a téglalapok területe az intervallumokbeli előfordulások relatív gyakoriságával (n_i/N), legyen arányos (1-2. ábra). Ez az ún. relatív gyakorisági hisztogram. Ha egyre több mérést végzünk és finomítjuk az osztályszélességet, az $f(x)$ valószínűség-sűrűségfüggvényt kapjuk, amelyet az ábrán folytonos vonal jelöl.

A sűrűségfüggvény értelmezése

Annak valószínűsége, hogy az x folytonos valószínűségi változó a és b közötti értéket vegyen föl (1-3. ábra):

$$P(a < x \le b) = \int_a^b f(x)dx \,. \qquad (1.4)$$



1-3. ábra. A folytonos valószínűségi változó sűrűségfüggvényének értelmezése

Mivel x folytonos valószínűségi változó, nincs értelme egy-egy érték valószínűségéről beszélni, ugyanis $P(x = x_0) = 0$ (bár ez nem lehetetlen esemény).

Az $f(x)$ sűrűségfüggvény tulajdonságai:

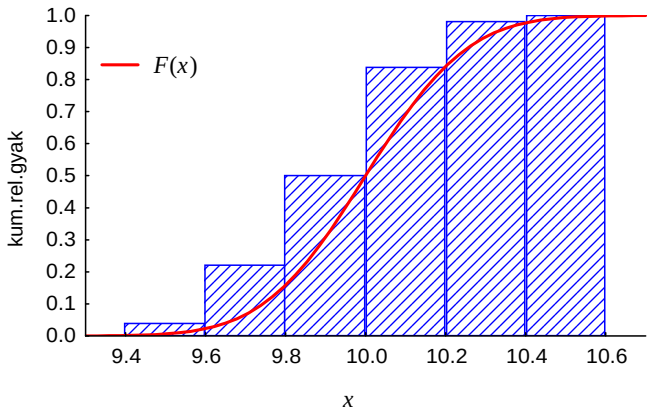
$f(x) \geq 0$ $\quad -\infty < x < \infty$, vagyis $f(x)$ értéke nem lehet negatív,

$\int_{-\infty}^{\infty} f(x)dx = 1$, vagyis az egész görbe alatti terület egységnyi.

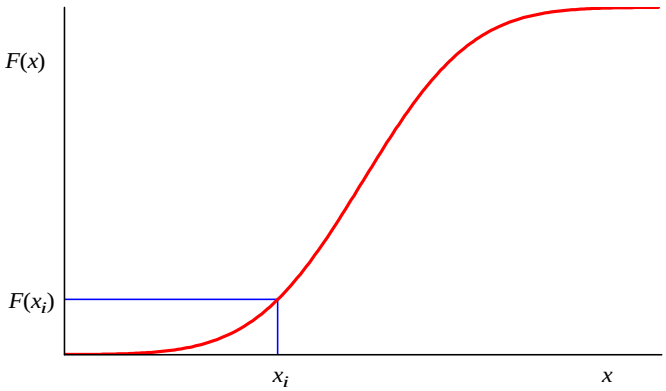
Ábrázoljuk a kumulált relatív gyakoriságokat (annak relatív gyakoriságát, hogy a valószínűségi változó x_i vagy annál kisebb értékeket vesz fel) x függvényében (1-4. ábra). Itt, ha egyre több mérést végzünk, az eloszlásfüggvényt kapjuk, ezért az előbbi kumulált relatív gyakorisági hisztogramot, ill. adatait tapasztalati eloszlásfüggvénynek is nevezik. Az eloszlásfüggvény a sűrűségfüggvény integrálja (l. az 1-5. ábrát):

$$F(x_i) = P(x \leq x_i) = \int_{-\infty}^{x_i} f(x)dx\,. \qquad (1.5)$$

A sűrűség-, ill. eloszlásfüggvény alakjának és paramétereinek ismerete jelenti a sokaság ismeretét.



1-4. ábra. Folytonos valószínűségi változó kumulált relatív gyakorisági hisztogramja és eloszlásfüggvénye



1-5. ábra. A folytonos valószínűségi változó eloszlásfüggvényének értelmezése

Paraméter és statisztika

A sokaságra vonatkozó valószínűség-sűrűség-, ill. -eloszlásfüggvény konstansai, ill. ezek származékai (momentumok stb.) a paraméterek. A méréssel (mintavétellel) ezek értékeiről, azaz a sűrűség- és eloszlásfüggvényről akarunk információt szerezni. A paraméterek analogonjai a minta jellemzői vagy más néven statisztikák. A paraméterek a sokaság tulajdonságai, míg a jellemzők (statisztikák) a mintáéi.

A legfontosabb paraméterek és statisztikák (jellemzők)

Várható érték

A várható érték definíciója folytonos valószínűségi változó esetén:

$$E(x) = \int_{-\infty}^{\infty} xf(x)dx = \mu, \tag{1.6}$$

ahol $f(x)$ a sűrűségfüggvény.

Diszkrét valószínűségi változóra:

$$E(x) = \sum_i x_i p(x_i). \tag{1.7}$$

A várható érték a sokaság tulajdonsága, tehát paraméter.
A mintára a várható értékkel analóg statisztika a számtani átlag:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i . \tag{1.8}$$

A valószínűségi változó függvényének várható értéke

Ha $\varphi(x)$ az x folytonos valószínűségi változó egyértékű valós függvénye, $\varphi(x)$ várható értékén a következő kifejezést értjük:

$$E[\varphi(x)] = \int_{-\infty}^{\infty} \varphi(x) f(x)dx . \tag{1.9}$$

Ennek alapján könnyen belátható, hogy $E(cx) = cE(x)$, és $E(c) = c$, ahol c konstans.

Ha x_1, x_2,..., x_n valószínűségi változók (pl. egy veendő minta elemei), a definícióból belátható, hogy

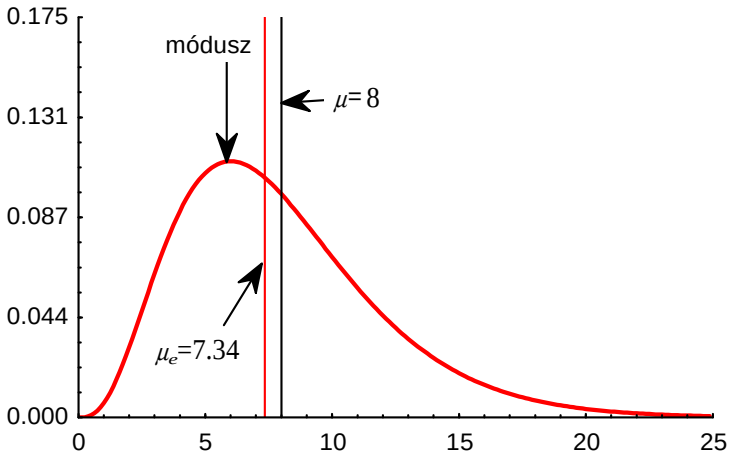
$$E(x_1 + x_2 + \cdots + x_n) = E(x_1) + E(x_2) + \cdots + E(x_n). \tag{1.10}$$

Medián

A medián az az érték, amelynél nagyobbat a valószínűségi változó ugyanolyan valószínűséggel vesz fel, mint kisebbet (1-6. ábra). A mediánt μ_e -vel jelölve ez a következőt jelenti:

$$F(\mu_e) = 0.5 . \tag{1.11}$$

A tapasztalati medián a nagyság szerint rendezett mintaelemek közül a középső. Páros mintaelemszám esetén a két középső érték számtani átlaga.



1-6. ábra. Módusz, medián, várható érték

Módusz

A módusz a valószínűségi változó legnagyobb valószínűségű értéke (a sűrűségfüggvény maximumhelye). Egy eloszlásnak több módusza is lehet.

A tapasztalati módusz a legnagyobb gyakoriságú osztály (a hisztogram legmagasabb téglalapjának) osztályindexe. Ha több móduszt találunk, általában több sokaság összekeveredésére gyanakodhatunk. Egycsúcsos szimmetrikus eloszlás esetében a módusz és a medián egybeesik a várható értékkel, aszimmetrikus esetben nem (1-6. ábra).

A variancia definíciója

Az x folytonos valószínűségi változóra:

$$Var(x) = \int_{-\infty}^{\infty} \left[x - E(x)\right]^2 f(x)dx = E\left[\left(x-\mu\right)^2\right] = \sigma^2 . \qquad (1.12)$$

Diszkrét valószínűségi változóra:

$$Var(x) = \sum \left[x_i - E(x)\right]^2 p\left(x_i\right) = E\left[\left(x-\mu\right)^2\right], \qquad (1.13)$$

azaz a várható értéktől való eltérés négyzetének várható értéke. Szokás (a magyar nyelvű szakirodalomban is) a következő jelölés: $D^2(x)$.

Megjegyzendő, hogy a magyar szakirodalomban a variancia helyett a szórásnégyzet elnevezést használják. Az elméleti és a tapasztalati szórásnégyzet megkülönböztetése végett tartottuk szükségesnek, hogy könyvünkben más kifejezést használjunk.

A variancia a sokaság tulajdonsága (ezért paraméter), a sűrűségfüggvény „szélességét” adja meg. A definíció alapján könnyen belátható, hogy

$$Var(cx) = c^2 Var(x), \qquad (1.14)$$

és független x_1 , x_2 ,.., x_n valószínűségi változókra (mint pl. egy minta elemeire):

$$Var\left(x_1 + x_2 + \cdots + x_n\right) = Var\left(x_1\right) + Var\left(x_2\right) + \cdots + Var\left(x_n\right). \qquad (1.15)$$

A variancia mintabeli analogonja a szórásnégyzet (más néven tapasztalati szórásnégyzet vagy korrigált tapasztalati szórásnégyzet):

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 . \tag{1.16}$$

1.2. A legfontosabb diszkrét eloszlások

Számos diszkrét eloszlás ismeretes, közülük számunkra most a binomiális és a Poisson-eloszlás a legfontosabbak.

A binomiális eloszlás

Dobjunk föl egy pénzérmét *n*-szer. Legyen *p* annak valószínűsége, hogy egy földobás eredménye fej legyen (ez hibátlan érménél 0.5). Annak valószínűségét, hogy a sorozatban éppen *x* legyen a fej dobások száma, a következő sűrűségfüggvény adja meg:

$$p(x) = \binom{n}{x} p^x \left(1-p\right)^{n-x} . \tag{1.17}$$

Általánosabban a binomiális eloszlás akkor használható, ha a vett minta eleme kétféle lehet. A gyártmány- vagy gyártásellenőrzésnél *p* a sokaságbeli (tételbeli) selejtarány, *x* az *n* elemű mintában talált selejtes darabok száma. Szükséges, hogy a mintavétel visszatevéssel történjék, vagyis a *k*-adik mintaelem ugyanolyan eséllyel legyen selejtes, mint a *k*+1-edik. Természetesen a gyakorlatban nem szokás a vett mintaelemeket visszatenni, ekkor a binomiális eloszlás csak közelítés, amely $n << N$ esetén teljesen jogos.

A binomiális eloszlású valószínűségi változó várható értéke és varianciája:

$$E(x) = np , \tag{1.18}$$

$$Var(x) = np\left(1-p\right) . \tag{1.19}$$

Ha a talált selejtes darabok száma helyett a mintabeli selejtarányt tekintjük valószínűségi változónak, ennek várható értéke és varianciája:

$$E\left(\frac{x}{n}\right) = p , \tag{1.20}$$

$$Var\left(\frac{x}{n}\right) = \frac{p\left(1-p\right)}{n} . \tag{1.21}$$

A mintabeli selejtarány is diszkrét valószínűségi változó, bár lehetséges értékei nem egész számok. Például 20 elemű mintában a talált selejtarány lehet 0, 1/20, 2/20 s.i.t.

A Poisson-eloszlás

Ritka események eloszlásának modellezésére használható, pl. a ritkán előforduló selejtes darabok tételenkénti száma, a műszakonkénti fonalszakadások száma, az üzemi

balesetek száma évente, a festési hibahelyek száma egy autón stb. A minőségbiztosításban elsősorban a termékegységen előforduló hibák eloszlásának modellezésére használják.

Annak feltételei, hogy a ritka esemény valamely idő-intervallumbeli, vagy adott egységbeli előfordulásainak száma Poisson-eloszlást kövessen:

a) bármely egységben bekövetkező eseménynek függetlennek kell lennie a többi egységbelitől;
b) az esemény bekövetkezésének valószínűsége bármely egységben azonos, és arányos az egység méretével;
c) annak valószínűsége, hogy két vagy több előfordulás következik be egy egységben, az egység méretének csökkentésével nullához tart.

Ha a binomiális eloszlásnál a p paraméter igen kicsi ($p \to 0$), a mintaelemszám pedig igen nagy ($n \to \infty$), de közben az $np = \lambda$ szorzat véges konstans ($\lambda > 0$), a valószínűségi változó Poisson-eloszlású lesz.

A Poisson-eloszlású valószínűségi változó sűrűségfüggvénye:

$$p(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}. \tag{1.22}$$

Várható értéke és varianciája:

$$E(x) = Var(x) = \lambda. \tag{1.23}$$

1.3. A legfontosabb folytonos eloszlás: normális eloszlás

A természetben akkor találkozunk normális eloszlással, ha sok, egymástól független, egyenként kis hatású tényező hatása összeadódik. Emiatt a közvetlenül mért, véletlenszerű ingadozásokat mutató adatok (tömeg, hőmérséklet stb.) jó közelítéssel normális vagy Gauss-féle eloszlású sokaságból vett mintának tekinthetők.

A Gauss-eloszlás sűrűség- és eloszlásfüggvénye:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \tag{1.24}$$

$$F(x_i) = \int_{-\infty}^{x_i} \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]dx. \tag{1.25}$$

A normális eloszlású valószínűségi változó várható értéke és varianciája:

$$E(x) = \mu, \tag{1.26}$$

$$Var(x) = \sigma^2. \tag{1.27}$$

A normális eloszlás szokásos rövid jelölése $N(\mu, \sigma^2)$, pl. $N(0, 1)$.

Ha az eloszlásfüggvény értékeit táblázatba akarnánk foglalni, háromdimenziós táblázatra lenne szükség, mivel $F(x)$ az x változón kívül a μ és σ paramétereket is tartalmazza. Célszerű tehát transzformációt keresnünk.

Normalizált (standardizált) normális eloszlás: u-eloszlás

Definiáljuk a következő valószínűségi változót:

$$u = \frac{x-\mu}{\sigma}. \tag{1.28}$$

Az új valószínűségi változó paraméterei:

$$E(u) = E\left(\frac{x-\mu}{\sigma}\right) = \frac{E(x)-\mu}{\sigma} = 0, \tag{1.29}$$

$$Var(u) = Var\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma^2} Var(x) = 1. \tag{1.30}$$

A két paraméter felhasználásával a normalizált (standardizált) normális eloszlás sűrűségfüggvénye:

$$f(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right). \tag{1.31}$$

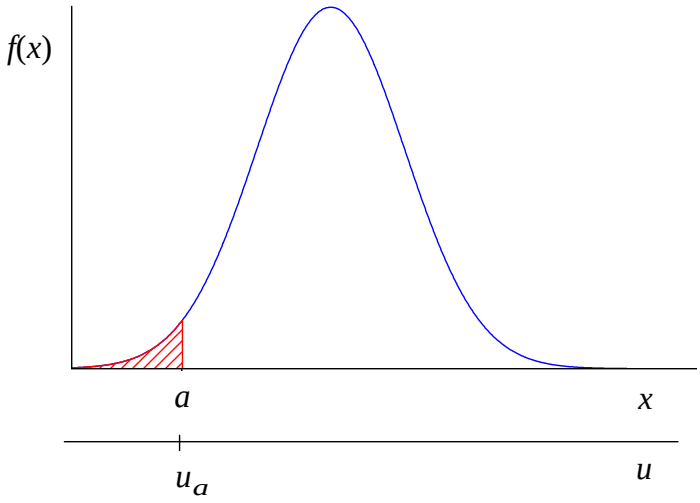
Minthogy a sűrűségfüggvényben egyetlen paraméter sem szerepel, a normalizált normális eloszlás eloszlásfüggvényének értékei kisméretű táblázatba foglalhatók (Függelék I. táblázat). E táblázat adatai bármilyen paraméterű normális eloszlásra használhatók a transzformációs képlet alkalmazásával.

Annak valószínűsége, hogy a $N(\mu, \sigma^2)$ eloszlású x valószínűségi változó nem haladja meg a értékét, a következő integrállal adható meg:

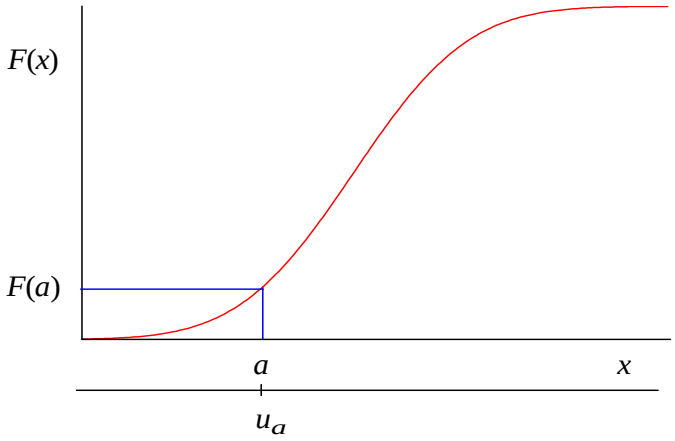
$$P(x \le a) = F(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx = \int_{-\infty}^{u_a} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du = F(u_a), \tag{1.32}$$

ahol $u_a = \dfrac{a-\mu}{\sigma}$.

A $P(x \le a)$ valószínűség értékét az 1-7*a*) ábrán a vonalkázott terület mutatja. A kettős vízszintes skála szemlélteti a transzformációt. Az 1-7*b*) ábra az eloszlásfüggvénnyel magyarázza ugyanezt.



1-7*a*) ábra. Standardizált normális eloszlású valószínűségi változó sűrűségfüggvénye



1-7*b*) ábra. Standardizált normális eloszlású valószínűségi változó és eloszlásfüggvénye

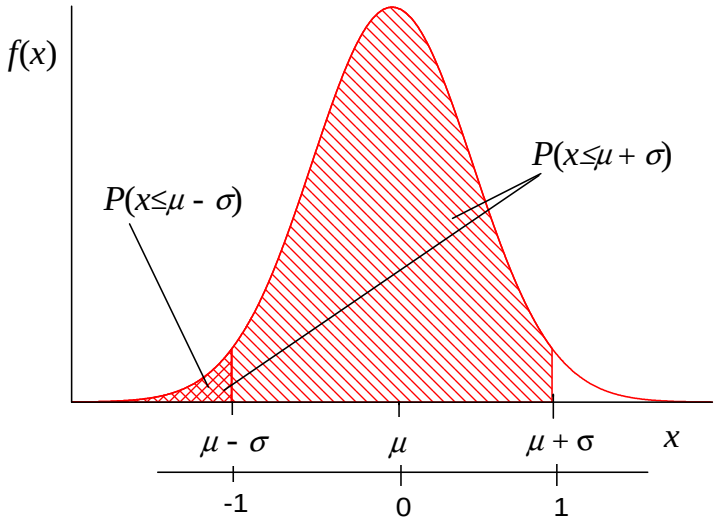
Tehát annak valószínűsége, hogy $x \le a$, megegyezik annak valószínűségével, hogy $u \le u_a = \dfrac{a-\mu}{\sigma}$.

1-1. példa

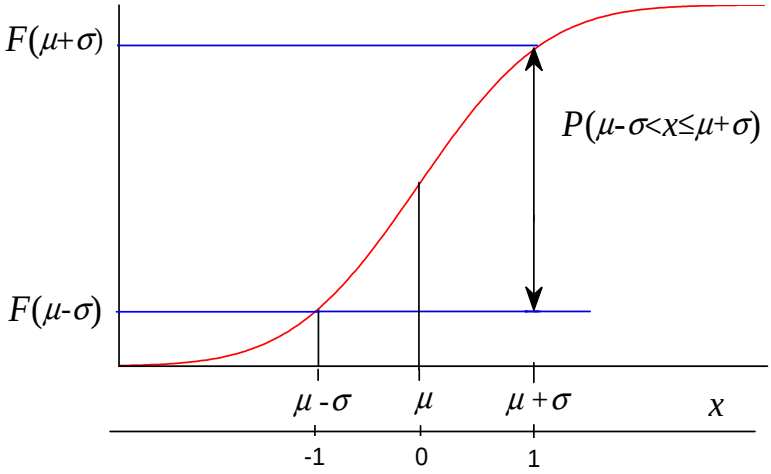
Határozzuk meg annak valószínűségét, hogy az x normális eloszlású valószínűségi változó a $(\mu-\sigma, \mu+\sigma)$ intervallumba eső értéket vesz fel!

$$P(\mu-\sigma < x \le \mu+\sigma) = F(\mu+\sigma) - F(\mu-\sigma)$$

Az összefüggést az 1-8*a*) és *b*) ábrák szemléltetik.



1-8*a*) ábra. A normális eloszlású valószínűségi változó $(\mu-\sigma,\mu+\sigma)$ intervallumbeli előfordulásának valószínűsége a sűrűségfüggvényen szemléltetve



1-8*b*) ábra. A normális eloszlású valószínűségi változó $(\mu-\sigma,\mu+\sigma)$ intervallumbeli előfordulásának valószínűsége az eloszlásfüggvényen szemléltetve

$$u_{\text{fölsõ}}=\frac{\mu+\sigma-\mu}{\sigma}=1 \qquad u_{\text{alsó}}=\frac{\mu-\sigma-\mu}{\sigma}=-1$$

A Függelék I. táblázatából $F(1)=0.84134$. Belátható, hogy mivel $f(x)$ szimmetrikus függvény és $F(\infty)=1$, $F(-a)=1-F(a)$. Így $F(-1)=1-F(1)=0.15866$.

$P(\mu-\sigma<x\le\mu+\sigma)=0.68268$; P ≈ 0.683, azaz a valószínűség 68.3 % .

Hasonló számítással adódik:

intervallum szélessége	$\pm\sigma$	$\pm2\sigma$	$\pm3\sigma$

P	0.68268	0.9545	0.9973

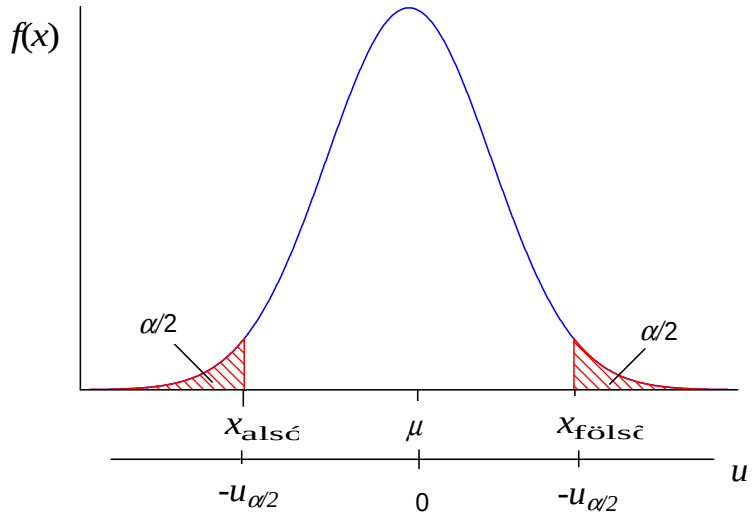
1-2. példa

Határozzuk meg, hogy egy $N(\mu,\sigma^2)$ normális eloszlású valószínűségi változó értékei milyen szimmetrikus intervallumban vannak 95 %-os, ill. 99 %-os valószínűséggel!

Határozzuk meg először az u normalizált normális eloszlású változó alsó és felső határértékét! Legyen α annak valószínűsége, hogy az érték az adott intervallumon kívül esik; szimmetrikus sűrűségfüggvényről lévén szó, $\alpha/2$ annak valószínűsége, hogy balra, ill. jobbra kiesik az intervallumból (1-9. ábra):

A Függelék I. táblázatából

α	0.05	0.01
1-α	0.95	0.99
1-$\alpha/2$	0.975	0.995
u	1.96	2.58



1-9. ábra. Az u-eloszlású valószínűségi változó $1-\alpha$ valószínűségű intervalluma

Térjünk vissza az eredeti x valószínűségi változóra és határozzuk meg a kérdéses intervallumot! Tehát $x_{\text{alsó}} = \mu - \sigma u_{\alpha/2}$; $x_{\text{fölsõ}} = \mu + \sigma u_{\alpha/2}$.

α	0.05	0.01
$x_{\text{alsó}}$	μ–1.96σ	μ–2.58σ
$x_{\text{fölső}}$	μ+1.96σ	μ+2.58σ