

Előkészítő jegyzetek

A valóságban és a gazdasági életben a legtöbb változó jövőbeli értéke ismeretlen.

Példák.: az államvizsga-jegyzetek, sikerül/nem sikerül az államvizsgátok, a jövedelmek egy év múlva, a holnapi lej-euró árfolyam.

Valószínűségi változó (vv): olyan változó, amelynek ismerjük a lehetséges értékeit és az ezekhez rendelt valószínűségeket, de a megvalósulása előtt nem ismerjük a pontos értékét.

Egy valószínűségi változó **megvalósulása** egy olyan esemény, amely során felveszi egy lehetséges értékét.

Példák.

1. Dobókocka. Melyek a lehetséges értékek és ezek valószínűségei?
2. A holnapi lej-euró árfolyam. Melyek a lehetséges értékek és ezek valószínűségei?

A lehetséges értékek szerint a valószínűségi változók lehetnek diszkrét és folytonosak.

Egy **folytonos** valószínűségi változó egy vagy több intervallumból minden értéket felvehet. Pl.: a holnapi lej-euró árfolyam.

Egy **diszkrét** valószínűségi változó csak jól elhatárolt értékeket vehet fel. Pl.: a dobókocka.

Egy másik példa: Két munkahelyre pályázol, az egyiken 1500 lej, a másikon 1000 lej a fizetés. Az elsőre 0.3, a másodikra 0.4 valószínűséggel vesznek fel. Tehát, a fizetésed egy valószínűségi változó. Melyek a lehetséges értékek és ezek valószínűségei?

Ebben a kurzusban többnyire folytonos valószínűségi változókkal fogunk dolgozni.

Folytonos valószínűségi változók

Egy X folytonos valószínűségi változót egyértelműen meg lehet határozni a hozzárendelt eloszlásfüggvény (F) vagy a hozzárendelt sűrűségfüggvény (f) által.

Az **eloszlásfüggvény** minden x valós számra értelmezett, és az $F(x) = P(X < x)$ képlettel határozzuk meg, ahol $P(X < x)$ az $(X < x)$ esemény valószínűségét jelöli.

Megjegyzés. $P(X = x) = 0$, és ezért $P(X < x) = P(X \leq x)$ bármely X folytonos valószínűségi változóra és x valós számra. Hasonlóan:

$$\begin{aligned} P[a \leq X \leq b] &= P[a \leq X < b] = P[a < X \leq b] \\ &= P[a < X < b]. \end{aligned}$$

Tehát egy folytonos vv lehetséges értékeinek a valószínűségét az eloszlásfüggvény **intervallumokra** határozza meg.

Az eloszlásfüggvény csak 0 és 1 közötti értékeket vesz fel és növekvő. Ezért $F(-\infty) = 0$ és $F(+\infty) = 1$.

Egy intervallum valószínűsége: $P[a < X < b] = F(b) - F(a)$.

A **sűrűségfüggvény** az eloszlásfüggvény deriváltja: $f(x) = F'(x)$.

Mivel F növekvő, ezért f nem vehet fel negatív értékeket. Más fontos tulajdonságok:

$$\begin{aligned} \int_{-\infty}^{+\infty} f(x) dx &= 1, \\ P(a < X < b) &= \int_a^b f(x) dx. \end{aligned}$$

Várható érték és variancia

Legyen X folytonos valószínűségi változó.

X **várható értéke**:

$$E[X] = \mu = \int_{-\infty}^{\infty} x f(x) dx.$$

Tulajdonság. Ha $g(X)$ az X vv egy függvénye, akkor

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx.$$

Például, ha $g(X) = X^2$, akkor $E[X^2] = \int_{-\infty}^{\infty} x^2 f(x) dx$.

Az X valószínűségi változó **varianciája (szórásnégyzete)**:

$$\text{var}(X) = \sigma^2 = E[(X - \mu)^2] = E[X^2] - \mu^2.$$

A fenti tulajdonság alapján

$$\text{var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2.$$

Legyenek $X_1, X_2, \dots, X_n$ megfigyelések egy bizonyos eloszláshoz (pl. a lej-euró napi árfolyam az elmúlt n munkanapon).

Gyakorlati szempontból a várható érték az a szám amely körül csoportosulnak a megfigyelések.

A várható érték becslőfüggvénye a megfigyelések mintaátlagá:

$$\bar{X} \equiv \frac{X_1 + X_2 + \dots + X_n}{n}.$$

A variancia egy pozitív szám, amely megmutatja, hogy a megfigyelések mennyire szóródnak szét a várható érték körül: minél nagyobb a variancia, annál nagyobb mértékű a szétszóródás.

A variancia (egy) becslőfüggvénye:

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

Az egyenletes eloszlás (*The uniform distribution*)

Az egyenletes eloszlást egy $[a, b]$ intervallumon értelmezzük, sűrűségfüggvénye

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{ha } a \leq x \leq b \\ 0 & \text{másképp,} \end{cases}$$

eloszlásfüggvénye

$$F(x) = \begin{cases} 0 & \text{ha } x < a \\ \frac{x-a}{b-a} & \text{ha } a \leq x \leq b \\ 1 & \text{ha } x > b. \end{cases}$$

Az eloszlás neve ("egyenletes") onnan származik, hogy egy ilyen eloszlású X vvra, annak a valószínűsége, hogy X az $[a, b]$ bizonyos részintervallumában van, csak az intervallum hosszától függ, és nem a helyétől.

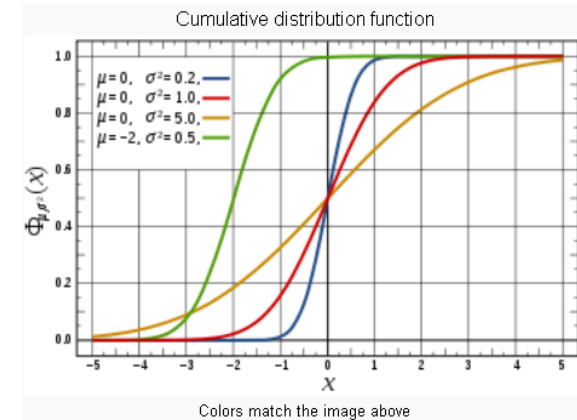
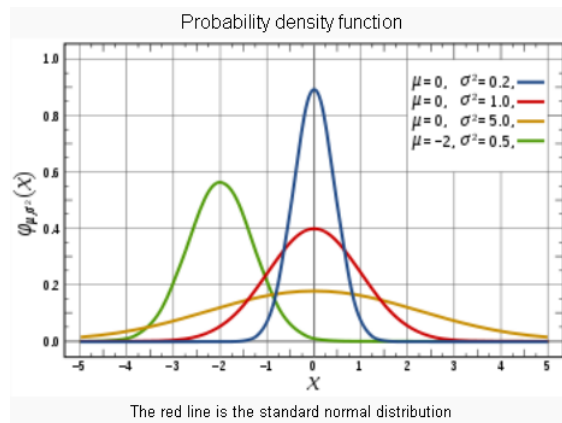
Várható érték: $\mu = \int_{-\infty}^{\infty} x f(x) dx = \frac{a+b}{2}.$

Variancia $\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2 = \frac{(b-a)^2}{12}.$

A normál eloszlás (*The normal distribution*)

A **normál** eloszlást egyértelműen meg lehet határozni a μ várható értéke és a σ^2 varianciája segítségével. Egy ilyen eloszlású X vv jelölése $X \sim N(\mu, \sigma)$. Sűrűségfüggvénye:

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right), \text{ ahol } -\infty < x < \infty.$$



Fennállnak az összefüggések:

$$\int_{-\infty}^{\infty} f(x|\mu, \sigma) dx = 1$$

$$E[X] = \int_{-\infty}^{\infty} x f(x|\mu, \sigma) dx = \mu$$

$$\text{var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x|\mu, \sigma) dx = \sigma^2.$$

A sűrűségfüggvény szimmetrikus a várható érték körül.

A normál eloszlás a fontosságát a Központi határeloszlás-tételből nyeri, lásd alább.

Ha $\mu = 0$ és $\sigma^2 = 1$ a **standard normál** eloszlást kapjuk, melynek sűrűségfüggvénye

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \text{ ahol } -\infty < z < \infty.$$

A standard normál eloszlás eloszlásfüggvénye

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

Ennek a megközelítő értékeit táblázatba foglalták; pontosabb értékeket statisztikai programok segítségével lehet kapni.

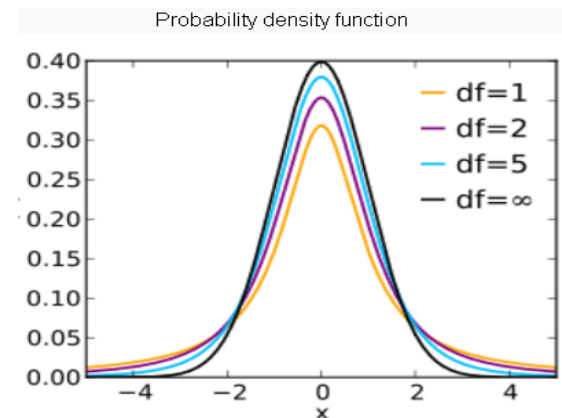
Egy nem standard normál eloszlású vvhoz rendelt valószínűségeket a következő tulajdonság segítségével számítjuk ki. Ha $X \sim N(\mu, \sigma^2)$, akkor

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

(számítsuk ki Z várható értékét és varianciáját).

A t eloszlás (*Student's t distribution*)

A normál eloszláshoz hasonlóan, a várható értéke körül szimmetrikus eloszlás.



A normál eloszlástól eltérően, a t eloszlás várható értéke mindig 0.

Egy paraméter határozza meg, amelyet **szabadsági fok**nak (*degree of freedom*) nevezünk (az ábrán df -fel jelölik).

Az eloszlás jelölése: $t(df)$.

Végtelen szabadsági fok esetén ($df = \infty$) a sűrűségfüggvénye megegyezik a standard normál eloszlás sűrűségfüggvényével (az ábrán a fekete vonal).

A t eloszlás bizonyos statisztikai tesztek miatt nagy fontosságú.

Példa: Megfigyelések átlagának a tesztelése

Tegyük fel, hogy az $X_1, X_2, \dots, X_n$ megfigyelések normál eloszlásúak, viszont nem tudjuk a várható értéküket és a varianciájukat. Azt szeretnénk tudni, hogy a megfigyelések 0 várható értékűek-e.

Statisztikai tesztek

Ehhez a statisztikában egy ún. **statisztikai tesztet** vagy **próbát** dolgoztak ki. A teszt logikai helyessége megkövetel néhány fogalmat és alapösszefüggést.

Egy statisztikai teszt legfontosabb összetevője a **tesztstatisztika** (vagy egyszerűen **statisztika**), ami a megfigyelésekből kiszámított olyan becslőfüggvény, amely alkalmas a probléma vizsgálatára. Ezt jelöljük T -vel.

Azt az állítást, amelyet vizsgálunk, **nullhipotézisnek** nevezzük, jele H_0 .

Egy, a nullhipotézist kizáró állítást **alternatív hipotézisnek** nevezünk, jele H_1 .

A **szignifikancia szint** egy α szám, amelyet leginkább 0.05-nek vesznek (még lehetséges értékek: $\alpha = 0.01$ vagy 0.1).

Az **elfogadási tartomány** az a (c_1, c_2) intervallum, amelyre

$$P(c_1 < T < c_2) = 1 - \alpha.$$

Ezt az intervallumot a T statisztika eloszlása alapján határozzuk meg.

A statisztikai teszt lépései:

1. megválasztjuk α -t,
2. meghatározzuk az elfogadási tartományt,
3. kiszámítjuk a T értékét a megfigyelésekből,
4. a teszt eredménye: ha $T \in (c_1, c_2)$, nem utasítjuk el H_0 -t; ha $T \notin (c_1, c_2)$, elutasítjuk H_0 -t.

A példa folytatása

$H_0 : \mu = 0$, $H_1 : \mu \neq 0$ (H_1 lehet $\mu > 0$ is, ebben az esetben az elfogadási tartományt másképp határozzuk meg). Legyen $\alpha = 0.05$. Az ehhez a nullhipotézishez kidolgozott statisztika:

$$T = \frac{\sqrt{n} \cdot \bar{X}}{s},$$

ahol

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

a szórás becslőfüggvénye.

A T statisztika egy vv, minek eloszlása: $T \sim t(n-1)$. Az elfogadási tartományt úgy határozzuk meg, hogy $c_1 = -c$, $c_2 = c$ és

$$P(-c < T < c) = 0.95.$$

A. Most tegyük fel, hogy van egy 100 megfigyelésből álló mintánk, amire $\bar{X} = -0.088$, $s = 0.874$. A teszt lépései:

1. $\alpha = 0.05$
2. $c = 1.96$, tehát az elfogadási tartomány $(-1.96, 1.96)$
3. $T = \frac{\sqrt{n} \cdot \bar{X}}{s} = \frac{10 \cdot (-0.088)}{0.874} = -1.007$
4. Mivel $T = -1.007 \in (-1.96, 1.96)$, nem utasítjuk el H_0 -t.

B. Most tegyük fel, hogy van egy másik 100 megfigyelésből álló mintánk, amire $\bar{X} = 0.774$, $s = 1.009$, és ugyanazt a nullhipotézis szeretnénk tesztelni. A teszt lépései:

1. $\alpha = 0.05$
2. $c = 1.96$, tehát az elfogadási tartomány $(-1.96, 1.96)$
3. $T = \frac{\sqrt{n} \cdot \bar{X}}{s} = \frac{10 \cdot 0.774}{1.009} = 7.671$
4. Mivel $T = 7.671 \notin (-1.96, 1.96)$, elutasítjuk H_0 -t.

Két vv közötti összefüggés

Az X_1, X_2 vvk kovarianciája:

$$\begin{aligned} cov(X_1, X_2) &= E[(X_1 - E[X_1])(X_2 - E[X_2])] \\ &= E[X_1 X_2] - E[X_1] E[X_2] \end{aligned}$$

azt fejezi ki, hogy a vvk mennyire nőnek vagy csökkennek együtt. Pl., ha $cov(X_1, X_2) > 0$, akkor X_1 és X_2 együtt nő vagy csökken. Ha $cov(X_1, X_2) < 0$, akkor X_1 növekedése X_2 csökkenésével történik. Ha $cov(X_1, X_2) \simeq 0$, akkor X_1 és X_2 változása között kevés összefüggés van.

Az X_1, X_2 korrelációs együtthatója:

$$corr(X_1, X_2) = \frac{cov(X_1, X_2)}{\sqrt{var(X_1) var(X_2)}}$$

hasonlóan méri a két vv változását, viszont az értéke le van szűkítve -1 és 1 közé. Ha $cov(X_1, X_2) = 0$ vagy $corr(X_1, X_2) = 0$ akkor azt mondjuk, hogy az X_1, X_2 vvk **nem korreláltak**.

Normál eloszlású vektorok

Az $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ kétváltozós v vektor normál eloszlású $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ várható értékkel és

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

varianciával (kovariancia mátrixszal) ha a sűrűségfüggvénye

$$f(x) = \frac{1}{2\pi |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)' \Sigma^{-1} (x - \mu) \right\},$$

ahol $(x - \mu)' = (x_1 - \mu_1, x_2 - \mu_2)$ az $x - \mu$ vektor transzponáltja (csak egy jelölés); tehát

$$(x - \mu)' \Sigma^{-1} (x - \mu) = (x_1 - \mu_1, x_2 - \mu_2) \Sigma^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix},$$

a $|\Sigma|$ a Σ mátrix determinánsa. Jelölés: $X \sim N(\mu, \Sigma)$.

Tulajdonságok:

- $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$ **peremeloszlások**
- $\text{cov}(X_1, X_2) = \sigma_{12}$
- Ha X_1, X_2 nem korreláltak, akkor $\sigma_{12} = 0$ és ezért

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}.$$

Ezt behelyettesítve az (X_1, X_2) sűrűségfüggvényébe, kapjuk, hogy

$$f(x_1, x_2) = f(x_1) \cdot f(x_2).$$

Ez az összefüggés azt jelenti, hogy az X_1, X_2 vvk **függetlenek**.

Tehát: Két normál eloszlású nem korrelált vv független.

Példa. A dobókocka dobás utáni értéke és a holnapi lej-euró árfolyam független vvk.

Hasznos tulajdonságok. Legyenek $X, Y, X_1, \dots, X_n$ nem feltétlenül normál eloszlású vvk.

1. $E[\alpha] = \alpha$, $\text{var}(\alpha) = 0$.
2. $E[\alpha X] = \alpha E[X]$.
3. $E[X + Y] = E[X] + E[Y]$.
4. $\text{Var}[\alpha X] = \alpha^2 \text{Var}[X]$.
5. $\text{Var}[X \pm Y] = \text{Var}[X] + \text{Var}[Y] \pm 2\text{Cov}[X, Y]$.
6. Ha X és Y nem korreláltak, akkor $E[XY] = E[X] E[Y]$ és $\text{Var}[X - Y] = \text{Var}[X] + \text{Var}[Y]$.
7. $\text{Var}[X_1 + \dots + X_n] = \text{Var}[X_1] + \dots + \text{Var}[X_n] + 2 \sum_{i < j} \text{Cov}[X_i, X_j]$.
8. Ha $X_1, \dots, X_n$ páronként nem korreláltak, akkor $\text{Var}[X_1 + \dots + X_n] = \text{Var}[X_1] + \dots + \text{Var}[X_n]$.
9. Ha $X \sim N(\mu, \sigma^2)$, akkor $\frac{X - \mu}{\sigma} \sim N(0, 1)$ mert $E\left[\frac{X - \mu}{\sigma}\right] = 0$ és $\text{var}\left(\frac{X - \mu}{\sigma}\right) = 1$.

A mintaátlag tulajdonságai nagy minták esetén

Legyenek $X_1, \dots, X_n$ független vvk μ várható értékkel és σ^2 variáciával. Az

$$\bar{X} = \frac{X_1 + \dots + X_n}{n}$$

mintaátlagra fennállnak az alábbiak:

$$E(\bar{X}) = E\left(\sum_{i=1}^n \frac{1}{n} X_i\right) = \sum_{i=1}^n \frac{1}{n} E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$
$$\text{var}(\bar{X}) = \sum_{i=1}^n \frac{1}{n^2} \text{var}(X_i) = \frac{1}{n^2} \sum_{i=1}^n \text{var}(X_i) = \frac{\sigma^2}{n}.$$

A nagy számok törvénye: Nagy minta esetén ($n \gg$) a mintaátlag jól megközelíti a valódi átlagot, vagyis $\bar{X} \simeq \mu$.

Ez abból is látszik, hogy $E(\bar{X}) = \mu$ és $\text{var}(\bar{X}) = \frac{\sigma^2}{n} \simeq 0$.

Központi határeloszlás-tétel: Nagy minta esetén ($n \gg$) a mintaátlag jól megközelíti az $N\left(\mu, \frac{\sigma^2}{n}\right)$ eloszlást.

Tehát nagy minta esetén a mintaátlag megközelítőleg normál eloszlású.