9. fejezet

Valószínűségszámítási és statisztikai alapok

Diszkrét és folytonos valószínűségi változók. Nagy számok törvénye, centrális határeloszlás tétele. Statisztikai becslések, klasszikus statisztikai próbák.

9.1. Valószínűségszámítási alapfogalmak

9.2. Valószínűségi változók

9.2.1. Definíció (Valószínűségi változó). *Valószínűségi változónak nevezünk egy eseménytér elemeihez valós számokat rendelő $X : Q \to \mathbb{R}$ függvényt.*

9.2.1. Diszkrét valószínűségi változók

9.2.2. Definíció (Diszkrét valószínűségi változó). *Az X valószínűségi változót diszkrétnek nevezzük, ha lehetséges értékeinek száma megszámlálható (véges vagy végtelen). Ezt röviden $X(Q)$-val jelöljük.*

9.2.3. Definíció (Diszkrét valószínűségeloszlás). *A $v : X(Q) \to \mathbb{R}$, $v(x_i) = p_i = P(X = x_i)$ függvényt az X változó eloszlásának nevezzük.*

9.2.4. Tétel (Az eloszlás tulajdonságai).

- $v(x_i) \geq 0$,
- $\sum_{i=1}^{n} v(x_i) = 1$.

9.2.5. Definíció (Diszkrét eloszlásfüggvény). *A fenti jelölésekkel* $F(x) = \sum_{x_i<x} p_i$ *függvényt az* X *diszkrét eloszlásfüggvényének hívjuk.*

9.2.6. Definíció (Várható érték). *Az előző jelölésekkel* X *valószínűségi változó várható értéke* $M(X) = \sum_{i=1}^{n} x_i v(x_i)$.

9.2.7. Megjegyzés. *Az* X *végtelen sokféle értéke esetén a várható érték csak akkor értelmes, ha a* $\sum_{i=1}^{\infty} v(x_i) x_i$ *sor abszolút konvergens.*

9.2.8. Definíció (Szórásnégyzet, szórás). *Az* X *valószínűségi változó szórásnégyzetének az* $(X-m)^2$ *valószínűségi változó várható értékét nevezzük:* $D^2(X) = \sum_{i=1}^{n} (x_i - m)^2 v(x_i)$.

A szórásnégyzet gyöke a szórás (D).

9.2.9. Tétel.

$$D^2(X) = \sum_{i=1}^{n} x_i^2 v(x_i) - m^2 = M(X^2) - m^2.$$

9.2.10. Megjegyzés. *Az* X *változó szórásának és várható értékének hányadosát relatív szórásnak is nevezzük.*

9.2.11. Tétel (A várható érték és a szórás tulajdonságai).

- $M(aX + b) = aM(X) + b$,
- $M(aX + bY) = aM(X) + bM(Y)$,
- $D^2(aX + b) = a^2 D^2(X)$.

9.2.12. Definíció (Normált valószínűségi változó).

$$X^* = \frac{X - M(X)}{D(X)}$$

Ekkor $M(X^*) = 0$ *és* $D^2(X^*) = 1$.

Együttes eloszlások, peremeloszlások

9.2.13. Definíció (Együttes eloszlás). *Legyenek X és Y valószínűségi változók. Ekkor az $X(Q) \times Y(Q)$ halmazon értelmezett*

$$w(x_i, y_j) = P(X = x_i, Y = y_j)$$

függvényt a két változó (vagy az $(X;Y)$ vektorváltozó) együttes eloszlásának hívjuk.

9.2.14. Definíció (Együttes várható érték). *Mivel két változó együttes eloszlása kielégíti az eloszlásra megfogalmazott feltételeket, képezhetjük belőle a két változó*

$$M(X;Y) = \sum_{i,j} x_i y_j w(x_i, y_j)$$

együttes várható értékét.

9.2.15. Definíció (Peremeloszlás). *X és Y együttes eloszlásából a*

$$v(x_i) = \sum_{j=1}^{m} w(x_i, y_j), \text{ illetve } u(y_j) = \sum_{i=1}^{n} w(x_i, y_j)$$

eloszlásokat az X, illetve Y változókra vonatkozó peremeloszlásoknak nevezzük.

9.2.16. Definíció (Kovariancia). *Két valószínűségi változó kovarianciája a $Cov(X, Y) = M((X - M(X))(Y - M(Y)))$ várható érték.*

9.2.17. Tétel. $Cov(X, Y) = M(XY) - M(X)M(Y)$

9.2.18. Definíció (Korrelációs együttható). *Ha X és Y valószínűségi változóknak létezik szórása és kovarianciája, akkor korrelációjuk az*

$$R(X, Y) = \frac{Cov(X, Y)}{D(X)D(Y)}.$$

A korrelációs együttható a valószínűségi változók közötti kapcsolat erősségét jellemzi.

9.2.19. Tétel (A korrelációs együttható tulajdonságai).

- *$R(X, Y) = 0$, ha X és Y függetlenek (nem megfordítható);*

- $R(X, aX + b) = 1$, *ha* $a > 0$, *mert* $Cov(X, aX + b) = aD^2(X)$;
- $|R(X, Y)| \leq 1$ *és* $|R(X, Y)| = 1 \Leftrightarrow X = aY + b$ *egyenlőség teljesül 1 valószínűséggel* ($a \neq 0, b \in \mathbb{R}$).

9.2.20. Tétel (Lineáris függvénykapcsolat). *Ha két valószínűségi változóra* $|R(X, Y)| = 1$, *akkor a két változó között függvénykapcsolat áll fenn, azaz* $\exists a, b \in \mathbb{R} : Y = aX + b$.

9.2.21. Definíció (Független valószínűségi változók). *Két valószínűségi változó független, ha* $P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(y = y_j)$.

Ebben az esetben az együttes eloszlás minden tagja a megfelelő peremeloszlások szorzataként áll elő, azaz: $w(x_i, y_j) = v(x_i) \cdot u(y_j)$.

9.2.22. Tétel (Független valószínűségi változók tulajdonságai). *Ha* X *és* Y *független valószínűségi változók, akkor:*

- $M(X + Y) = M(X) + M(Y)$,
- $D^2(X + Y) = D^2(X) + D^2(Y)$.

9.2.2. Folytonos valószínűségi változók

9.2.23. Definíció (Eloszlásfüggvény). *Az* $x \in \mathbb{R} : F(x) = P(X < x)$ *függvényt az* X *valószínűségi változó eloszlásfüggvényének nevezzük.*

9.2.24. Tétel (Az eloszlásfüggvény tulajdonságai).

1. $F(x) \geq 0$,
2. F *monoton növekedő,*
3. $lim_{x \to -\infty} F(x) = 0$ *és* $lim_{x \to +\infty} F(x) = 1$,
4. F *minden pontjában balról folytonos.*

9.2.25. Definíció (Folytonos valószínűségi változó). *Ha egy valószínűségi változó eloszlásfüggvénye folytonos, akkor eloszlását is folytonosnak (abszolút folytonosnak) nevezzük.*

9.2.26. Definíció (Sűrűségfüggvény). *Az* X *folytonos valószínűségi változó sűrűségfüggvénye* f, *ha* $F(X) = \int_{-\infty}^{x} f(t) \mathrm{d}t$ *értelmes.*

9.2.27. Tétel.

- *A sűrűségfüggvény nemnegatív és* $\int_{-\infty}^{\infty} f(t)\mathrm{d}t = 1$,
- $P(a \leq x < b) = F(b) - F(a)$,
- $P(a \leq x \leq b) = \int_a^b f(x)\mathrm{d}x$.

9.2.28. Definíció (Várható érték). *AZ X abszolút folytonos valószínűségi változó várható értéke* $M(X) = \int_{-\infty}^{\infty} xf(x)\mathrm{d}x$, *feltéve, hogy ez az improprius integrál abszolút konvergens.*

9.2.29. Definíció (Szórásnégyzet, szórás). *Ha X folytonos eloszlású valószínűségi változó sűrűségfüggvénye* $f(x)$, *akkor szórásnégyzete*

$$D^2(X) = M((X - M(X))^2) = \int_{-\infty}^{\infty} (x - M(X))^2 f(x)\mathrm{d}x,$$

szórása a szórásnégyzet négyzetgyöke (feltéve, hogy az integrál létezik).

9.2.30. Tétel. *A szórásnégyzet felírható* $D^2 = M(X^2) - M^2(X)$ *alakban.*

9.2.31. Megjegyzés. *A diszkrét valószínűségi változók várható értékre és szórásra vonatkozó összefüggései igazak folytonos valószínűségi változókra is.*

9.2.3. Tételek

9.2.32. Tétel (A centrális határeloszlás tétele). *Ha* $X_1, \ldots, X_n$ *azonos eloszlású, független, véges várható értékű és szórású valószínűségi változók, akkor*

$$lim_{n\to\infty} P\left(\frac{X_1 + \ldots + X_n - nm}{\sigma\sqrt{n}} < x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{t^2}{2}}\mathrm{d}t = \Phi(x),$$

ahol $m = M(X_k)$, $\sigma = D(X_k)(k = 1, 2, \ldots)$, nm *az* $X_1 + \ldots + X_n$ *összeg várható értéke,* $\sigma\sqrt{n}$ *a szórása, és* $\Phi(x)$ *a standard normális eloszlásfüggvény.*

Azaz sok független valószínűségi változó összege normális eloszlású.

9.2.33. Tétel (A nagy számok törvénye). *Legyenek* $X_1, \ldots, X_n$ *azonos eloszlású, független, azonos (véges) várható értékű (m) és szórású valószínűségi változók. Ekkor az* $\bar{X_n} = \frac{X_1 + \ldots + X_n}{n}$ *változóra:*

$$lim_{n\to\infty} P\left(\left|\bar{X_n} - m\right| \geq \varepsilon\right) = 0,$$

azaz a számtani közép sztochasztikusan konvergál a várható értékhez.

9.2.34. Definíció (Sztochasztikus konvergencia). *Valószínűségi változók ξ_n, $n = 1, 2, \ldots$, sorozata akkor konvergál sztochasztikusan egy ξ valószínűségi változóhoz, ha (egyrészt ezek a valószínűségi változók ugyanazon az (Ω, A, P) valószínűségi mezőn vannak definiálva, másrészt) minden $\varepsilon > 0$ számra:*

$$lim_{n\to\infty} P\left(|\xi_n(\omega) - \xi(\omega)| > \varepsilon\right) = 0.$$

9.3. Statisztikai becslések

A statisztikai becslések feladata, hogy ha sejtjük egy minta eloszlását, közelíthessük az eloszlás ismeretlen paramétereit.

9.3.1. Definíció (Statisztika). *Egy ismeretlen a paramétert közelítő mintaelemekből képzett $b(X_1, X_2, \ldots, X_n)$ függvényt az a paraméter statisztikai becslésének (röviden statisztikának) nevezzük.*

9.3.2. Definíció (Torzítatlanság). *Azt mondjuk, hogy a b statisztika torzítatlanul becsli az a paramétert, ha várható értéke $M(b(X_1, \ldots, X_n)) = a$.*

9.3.3. Definíció (Konzisztencia). *Egy statisztikai becslést konzisztensnek nevezünk, ha elegendően nagy mintaelemszámra tetszőlegesen közelíti a becsült paramétert, azaz $\forall\varepsilon, \delta \in \mathbb{R} : \exists N \in \mathbb{N} :$ ha $n > N$, akkor*

$$P\left(|b(X_1, \ldots, X_n) - a| \geq \varepsilon\right) \leq \delta.$$

9.3.4. Tétel (A várható érték becslése). *A várható értéknek minden eloszlás esetén torzítatlan becslését adja a tapasztalati (empirikus) várható érték, azaz a mintaelemek átlaga:*

$$M(X) \approx \frac{X_1 + \ldots X_n}{n}.$$

9.3.5. Tétel (A szórás becslése). *A szórásnégyzetnek minden eloszlás esetén torzítatlan becslését adja a korrigált tapasztalati szórásnégyzet.*

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

9.3.1. Maximum-likelihood becslés

Tegyük fel, hogy $X_1, \ldots, X_n$ adott minta, mellyel az $a$paramétert szeretnénk becsülni. Ekkor a minta közös sűrűségfüggvénye

$$f_a(x) = \prod_{i=1}^{n} f_a(x_i).$$

Ekkor a maximum-likelihood becslésének nevezzük azt az $\hat{a}(X_1, \ldots, X_n)$ függvényt, melyre $f_{\hat{a}}(x)$ maximális.

A feladat tehát az $L = \prod_{i=1}^{n} f_a(x_i)$ *likelihood-függvény* szélsőértékeinek keresése. Egy függvény és logaritmusa ugyanott veszik fel szélsőértékeiket, melyeket általánosan deriválás segítségével határozhatunk meg, tehát a megoldandó feladat:

$$\frac{\mathrm{d} \ln L}{\mathrm{d} a} = 0.$$

9.3.6. Megjegyzés.

- *Diszkrét eloszlás esetén is alkalmazható a módszer, ekkor a likelihood-függvény:*

  $$L = \prod_{i=1}^{m} p_a(x_i)^{r_i},$$

  ahol r_i at i-edik mintaelem gyakorisága.

- *Többváltozós függvények deriválásával szükség esetén egyszerre több paraméter is becsülhető.*

- *A gyakorlat szempontjából elegendően általános feltételek mellett megmutatható, hogy a maximum-likelihood becslés konzisztens és minimálishoz közeli szórású.*

9.3.2. Konfidenciaintervallumok

Ez a módszer nem konkét értéket határoz meg a paraméter becsléseként, hanem adott valószínűséghez rendel intervallumot, melyben a paraméter a megadott valószínűséggel található.

9.3.7. Megjegyzés. *A konfidenciaintervallum hossza a mintaelemszám növelésével annak négyzetgyökével arányosan csökken.*

Várható érték becslése. Legyen X normális eloszlású változó ismert σ szórással. Ekkor várható értékének $1-p$ megbízhatóságú konfidenciaintervalluma:

$$\left[X - u_p \frac{\sigma}{\sqrt{n}}; X + u_p \frac{\sigma}{\sqrt{n}}\right],$$

ahol

$$\Phi(u_p) = 1 - \frac{p}{2}.$$

Ha megköveteljük, hogy a konfidenciaintervallum félhossza legfeljebb d lehessen, akkor a mintaelemszámra az

$$n \geq u_p^2 \frac{\sigma^2}{d^2}$$

alsó korlát adódik.

Szórás becslése. Keressük az $N(m, \sigma)$ ismert várható értékű normális eloszlás szórását. Jelölje s^{*^2} a minta tapasztalati szórásnégyzetét! Ekkor a σ^2 szórásnégyzet $1-p$ megbízhatóságú konfidenciaintervalluma:

$$\left[n \cdot \frac{s^{*^2}}{\chi^2_{\frac{p}{2}}}; n \cdot \frac{s^{*^2}}{\chi^2_{1-\frac{p}{2}}}\right],$$

ahol χ^2_p az n szabadságfokú χ^2 eloszlás p valószínűséghez és n szabadságfokhoz tartozó értéke.

9.4. Statisztikai próbák

A statisztikai próbák feladata általában valamely mintá(k)ra vonatkozó feltevésünk megerősítése vagy elutasítása.

9.4.1. Alapfogalmak

A statisztikai próba alapja mindig egy statisztika, azaz egy $h(X_1, \ldots, X_n)$ függvény, amire vonatkozóan a hipotéziseinket felállítottuk.

Nullihpotézis, ellenhipotézis. *Alaphipotézis*nek vagy *nullhipotézis*nek (H_0) nevezzük az alapfeltevésünket, amit igazolni szeretnénk. Ezzel szemben áll az *ellenhipotézis*, ami akkor tejesül, ha a nullhipotézis nem igaz.

Elfogadási tartomány, kritikus tartomány. *Elfogadási tartomány*nak nevezünk egy olyan halmazt, amelyben a nullhipotézis teljesülése esetén a vizsgált statisztika értéke nagy valószínűséggel $(1-p)$ elhelyezkedik. Ezzel szemben a *kritikus tartomány*ba várhatóan akkor esik a h statisztika értéke, ha az ellenhipotézis teljesül.

Hibák.

	H_0 igaz	H_0 hamis
H_0 elfogadása	helyes döntés	másodfajú hiba
H_0 elvetése	elsőfajú hiba	helyes döntés

9.4.1. Definíció (Terjedelem). *Egy próba terjedelme az elsőfajú hiba valószínűségének felső határa.*

9.4.2. Definíció (Erőfüggvény). *Ha β a másodfajú hiba valószínűsége, akkor a próba erőfüggvénye $1-\beta$.*

9.4.3. Definíció. *Azonos terjedelmű próbák közül erősebbnek nevezünk azt, amelyiknek az erőfüggvénye minden ponton nem kisebb a másikénál.*

9.4.2. u-próbák

Egymintás eset

Adott egy X normális eloszlású, ismert szórású változó n elemű mintája. Hipotézisünk a várható értékre vonatkozik, azaz $H_0 : m = m_0$. Legyen a próbastatisztika:

$$u = \sqrt{n}\frac{\bar{X} - m_0}{\sigma}.$$

Ekkor a p valószínűség mellett elfogadjuk a nullhipotézist, ha $|u| \leq u_{\frac{p}{2}}$, ahol $u_{\frac{p}{2}}$ a standard normális eloszlás $1 - \frac{p}{2}$ kvantilise.

Ha az ellenhipotézis egyoldalú, akkor az elfogadási tartomány $u \leq u_p$-re, illetve $u \geq -u_p$-re módosul.

Kétmintás eset

Két független (a fenti feltételeket teljesítő) mintára $(X,\ Y)$ vonatkozó nullhipotézisünk, hogy várható értékük egyenlő. Erre a próbastatisztika:

$$u = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}.$$

9.4.4. Tétel. *Az u-próbák konzisztens, torzítatlan és legerősebb próbák.*

9.4.3. t-próba

Egymintás eset

Ha a vizsgált normális eloszlás szórása sem ismert, de feltehető, hogy azonos, akkor az u-próba helyett t-próbát alkalmazunk. Ennek próbastatisztikája csak annyiban különbözik, hogy a szórás helyett a korrigált tapasztalati szórásnégyzettel dolgozunk, azaz:

$$t = \sqrt{n}\frac{\bar{X} - m_0}{s^{*2}}.$$

Az elfogadási tartomány ekkor: $|t| \leq t_{\frac{p}{2},n-1}$, ahol a jobb oldalon a $\frac{p}{2}$ höz tartozó, $n-1$ szabadságfokú *Student*-eloszlás áll.

Egyoldali ellenhipotézis esetén az u-próbához hasonlóan módosul a tartomány.

Kétmintás eset

Adott X és Y, a fenti feltételeket teljesítő mintákra vizsgáljuk azt a nullhipotézist, mely szerint a két változó várható értéke azonos. A próbastatisztika:

$$t = \sqrt{\frac{nm(n+m-2)}{n+m}}\frac{\bar{X} - \bar{Y}}{\sqrt{s_x^{*2} + s_y^{*2}}}.$$

9.4.5. Tétel. *A t-próba legerősebb próba.*

9.4.4. F-próba

Ha két, normális eloszlású, ismeretlen paraméterekkel rendelkező mintáról szeretnénk eldönteni, hogy szórásuk azonos-e, akkor F-próbát alkalmazunk. Ennek próbastatisztikája:

$$F = \max\left(\frac{s_x^{*2}}{s_y^{*2}}, \frac{s_y^{*2}}{s_x^{*2}}\right).$$

A elfogadási tartomány határa az $(n-1, m-1)$ szabadságfokú F-eloszlás $1 - \frac{p}{2}$ kvantilise, ahol n a számlálóbeli, m a nevezőbeli minta elemszáma.

9.4.5. Illeszkedésvizsgálat, homogenitásvizsgálat

Az *illeszkedésvizsgálat* ún. nemparaméteres próba, mellyel azt vizsgáljuk, hogy egy minta a megadott eloszlásból származhat-e.

Egy gyakran használt próba a *Kolmogorov-Szmirnov próba*, mely a tapasztalati eloszlásfüggvények eltérésének maximumán alapul.

A *homogenitásvizsgálat* feladata két minta alapján megvizsgálni, hogy azok azonos eloszlásból származnak-e.

(?) Mindkét módszernél támaszkodunk a minta osztályozására és az ez alapján képzett tapasztalati eloszlásfüggvényre.