Segédanyag a Valószínűségszámítás és statisztika tantárgyhoz

2011. november 30.

Definíció. **Véges valószínűségi mező:** $(\Omega, \mathcal{A}, P)$ hármas, ahol:

- $\{\omega_1, ..., \omega_n\} = \Omega$: eseménytér; elemei: elemi események
- $\{A, B, ...\} = \mathcal{A} \subset 2^\Omega$, ahol $A, B, ...$: események
- P: valószínűségi mérték
  - $P(\Omega) = 1$
  - $P(A) \geq 0$ minden A eseményre
  - $P(A \cup B) = P(A) + P(B)$ minden A és B egymást kizáró eseményre $(A \cap B = \emptyset)$.

Definíció. **Kolgomorov-féle vsz.-i mező:** $(\Omega, \mathcal{A}, P)$ hármas, ahol:

- Ω: nemüres halmaz
- $\mathcal{A} \subset 2^\Omega$ σ-algebra
- P: $\mathcal{A} \to [0; 1]$ halmazfüggény, amelyre
  - $P(\Omega) = 1$
  - $P(A) \geq 0 \quad \forall A \in \mathcal{A}$-ra
  - páronként kizáró $A_1, A_2, ... \in \mathcal{A}$ eseményekre
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
$P\left(\overline{A}\right) = 1 - P(A)$
Permutáció: n elem összes lehetséges sorrendje: $n!$

Ismétléses permutáció: n elem összes lehetséges sorrendje, ha ezek közül $k_1, ..., k_m$ darab megegyezik: $\frac{n!}{k_1! \cdot ... \cdot k_m!}$

Adott n elem, ebből k darabot kiveszünk

	Kombináció: a kihúzás sorrendje NEM számít (nem számozottak az elemek)	**Variáció**: a kihúzás sorrendje számít (számozottak az elemek)
Visszatevés nélkül (ismétlés nélkül)	$\binom{n}{k} = \frac{n!}{k!(n-k)!}$	$\frac{n!}{(n-k)!}$
Visszatevéssel (ismétléssel)	$\binom{n+k-1}{k}$	n^k

Mintavétel: Adott N termék, ezek közül M selejtes. Az összes termékből kiveszünk n darabot. Mi a valószínűsége, hogy ezek között k selejtes lesz? $(k = 0, 1, ..., n)$

- Visszetevés nélkül: $\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$
- Visszatevéssel: $\binom{n}{k} p^k (1-p)^{n-k}$ ahol $p = \frac{M}{N}$ a selejtarány

---

Feltételes valószínűség: Ha B bekövetkezett, mi a valószínűsége, hogy A bekövetkezik?
$P(A|B) = \frac{P(A \cap B)}{P(B)}$ $(P(B) \neq 0)$

Definíció. **Teljes eseményrendszer (TER)** $B_1, B_2, ...$ események teljes eseményrendszert alkotnak, ha

- $B_i \cap B_j = \emptyset$ minden i$\neq$j-re
- $\bigcup_{i=1}^{\infty} B_i = \Omega$

Tétel. **Teljes valószínűség tétele:** Legyen $B_1, B_2, ...$ teljes eseményrendszer, A tetszőleges esemény, $P(B_j) > 0$ minden j-re
Ekkor $P(A) = \sum_{j=1}^{\infty} P(A|B_j)P(B_j)$.

Tétel. **Bayes-tétel:** Legyen $B_1, ..., B_n$ teljes eseményrendszer, A tetszőleges esemény, $P(B_j) > 0$ minden j-re
Ekkor $P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{j=1}^{\infty} P(A|B_j)P(B_j)}$.

Definíció. **Események függetlensége:** A és B események függetlenek, ha $P(A \cap B) = P(A) \cdot P(B)$ (A esemény bekövetkezése nem befolyásolja B esemény bekövetkezését, és fordítva).

Definíció. **Valószínűségi változó:** X: $\Omega \to \mathbb{R}$ mérhető függvény, azaz amire $\{\omega : X(\omega) \in B\} \in \mathcal{A}$ minden B$\subseteq \mathbb{R}$ nyílt halmazra.

Definíció. **Valószínűségi változó eloszlása:** $Q_X(B) = P(X \in B) = P(\omega : X(\omega) \in B)$

Definíció. **Diszkrét valószínűségi változó:** értékkészlete legfeljebb

megszámlálhatóan végtelen, azaz $\{x_1, ..., x_n, ...\}$ elemekből áll.
Ekkor eloszlása: $p_i := P(X = x_i) = P(\omega : X(\omega) = x_i)$

Tétel. **Binomiális tétel:** $(a+b)^n = \sum\limits_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$.

Geometriai sor összege: $\sum\limits_{n=0}^{\infty} q^n = \frac{1}{1-q}$, ha |q|<1.
Konvergenciatartományon belül "be lehet deriválni" egy végtelen sort, így $\sum\limits_{n=1}^{\infty} nq^{n-1} = \frac{1}{(1-q)^2}$, ha |q|<1.

Minta: $X_1, ..., X_n$ valószínűségi változó sorozat. (Jel. $\mathbf{X} = X_1, ..., X_n$)
A továbbiakban feltesszük, hogy függetlenek és azonos eloszlásúak. Magyarosan rövidítve FAE minta, de gyakrabban használják az angol *i.i.d. minta* rövidítést (independent, identically distributed).
Az elméleti értékeket nagy, a konkrét, realizált mintából számolt értékeket mindig kis betű fogja jelölni, azaz minta esetén $x_1, ..., x_n$.

Statisztika: a minta valamely függvénye: $T : \mathbf{X} \to ...$
Becslés: a minta eloszlásának ismeretlen paraméterét közelíti a minta segítségével
Megj.: Minden becslés statisztika.

Néhány lényeges statisztika:

- *Rendezett minta*: $X_1^* \leq ... \leq X_n^*$ nem csökkenő sorrendbe tesszük a mintaelemeket
- *Mintaátlag*: $\overline{X} = \frac{\sum\limits_{i=1}^{n} X_i}{n}$
- *Tapasztalati szórás*: $S_n = \sqrt{\frac{\sum\limits_{i=1}^{n} (X_i - \overline{X})^2}{n}}$
  Értelmezése: az átlagtól való átlagos eltérés abszolút mértékegységben
- *Korrigált tapasztalati szórás*: $S_n^* = \sqrt{\frac{\sum\limits_{i=1}^{n} (X_i - \overline{X})^2}{n-1}}$
- *Szórási együttható*: $V = \frac{S_n}{\overline{X}}$
  Értelmezése: az átlagtól való átlagos eltérés százalékban
  Megj.: relatív szórásnak is hívják
- *Tapasztalati eloszlásfüggvény*: $F_n(x) = \frac{\sum\limits_{i=1}^{n} I(X_i < x)}{n}$
  ahol $I(X_i < x) = \begin{cases} 1 & \text{ha } X_i < x \\ 0 & \text{ha } X_i \geq x \end{cases} \rightsquigarrow$ karakterisztikus függvény
- *z-kvantilis*: $q_z = \inf\{x : F(x) > z\}$, és amennyiben F invertálható, akkor $q_z = F^{-1}(z)$-re egyszerűsödik
  Értelmezése: a mintaelemek z-ed része q_z-nél kisebb, $(1-z)$-ed része q_z-nél nagyobb
  Realizált mintából sokféleképpen számolható, interpolációs módszer:
  1.) Sorszám megállapítása: $(n+1)z = e + t$ (e:egészrész, t:törtrész)
  2.) $q_z = X_e^* + t(X_{e+1}^* - X_e^*)$
- *kvartilisek*: speciális kvantilisek
  - $Q_1 := q_{\frac{1}{4}} \rightsquigarrow$ alsó kvartilis
  - $Q_2 = Me := q_{\frac{1}{2}} \rightsquigarrow$ medián (középső mintaelem)
  - $Q_3 := q_{\frac{3}{4}} \rightsquigarrow$ felső kvartilis

---

Legyen $(\Omega, \mathcal{A}, P)$ valószínűségi mező, X: $\Omega \to \mathbb{R}$ valószínűségi változó.

Definíció. **Várható érték** X várható értéke: EX=$\int\limits_{\Omega} X \, dP$, ha ez létezik.

Definíció. **l. momentum** : EX^l=$\int\limits_{\Omega} X^l \, dP$, ha ez létezik.

Definíció. **X szórásnégyzete** : D^{2}X=E[(X-EX)]^2 = EX^2-E^{2}X.

Definíció. **X szórása** : DX=$\sqrt{D^2X}$.

Legyen X diszkrét valószínűségi változó, ami az x_1,x_2,... értékeket veszi fel, p_1,p_2,... valószínűségekkel. Ekkor EX=$\sum\limits_{i=1}^{\infty} x_k p_k$, ha a végtelen összeg abszolút konvergens. Ugyanígy EX^l=$\sum\limits_{i=1}^{\infty} (x_k)^l p_k$, ha a végtelen összeg abszolút konvergens.
Nevezetes diszkrét eloszlások:

Eloszlás neve	Jelölése	Eloszlása	EX	D^2X
Karakterisztikus (indikátorvált.)	Ind(p)	$P(X=1)=p$ $P(X=0)=1-p$	p	$p(1-p)$
Geometriai (Pascal)	Geo(p)	$P(X=k)=p(1-p)^{k-1}$ k=1,2,...	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Hipergeometriai	Hipgeo(N,M,n)	$P(X=k)=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$ k=0,1,...,n	$n\frac{M}{N}$	$n\frac{M}{N}\left(1-\frac{M}{N}\right)\left(1-\frac{n-1}{N-1}\right)$
Binomiális	Bin(n,p)	$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ k=0,1,...,n	np	$np(1-p)$
Negatív binomiális	NegBin(n,p)	$P(X=k)=\binom{k-1}{n-1}p^n(1-p)^{k-n}$ k=n,n+1,...	$\frac{n}{p}$	$\frac{n(1-p)}{p^2}$
Poisson	Poi(λ)	$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$ k=0,1,...	λ	λ

Előfordulásuk:

- Indikátor változó: egy p valószínűségű esemény bekövetkezik-e vagy sem
- Geometriai: hányadikra következik be először egy p valószínűségű esemény
- Hipergeometriai: visszatevés nélküli mintavétel
- Binomiális: visszatevéses mintavétel
- Negatív binomiális: hányadikra következik be n. alkalommal egy p valószínűségű esemény

Állítás. Legyenek $X, Y, X_1, ..., X_n$ valószínűségi változók; $c, c_i, a, b \in \mathbb{R}$. Ekkor

- $E(X+Y) = EX + EY$;
- $E(cX) = cEX$;
- $E\sum_{i=1}^{n} c_i X_i = \sum_{i=1}^{n} c_i EX_i$;
- $D^2(aX+b) = a^2 D^2 X$.

Definíció. **X val.változó eloszlásfüggvénye:** $F_X(x) = P(X < x)$.

Amennyiben egyértelmű, melyik val.változó eloszlásfüggvényéről van szó, F(x)-et írunk.

Állítás. Az eloszlásfüggvény tulajdonságai:

- $0 \leq F_X(x) \leq 1$;
- monoton növő;
- balról folytonos;
- $\lim\limits_{x\to-\infty} F(x) = 0$, $\lim\limits_{x\to\infty} F(x) = 1$.

Állítás. Tetszőleges X val.változó esetén

- $P(a \leq X < b) = F(b) - F(a)$;
- $P(a < X \leq b) = F(b+) - F(a+)$.

Definíció. **X val.változó abszolút folytonos**, ha létezik olyan f(x) függvény, amelyre $F(x) = \int\limits_{-\infty}^{x} f(t)\,dt$. Ilyenkor f(x)-et **sűrűségfüggvény**nek hívjuk.

Állítás. Legyen X abszolút folytonos eloszlású. Ekkor

- f(x)=F'(x);
- f(x) ≥ 0;
- $\int\limits_{-\infty}^{\infty} f(x)\,dx = 1$;
- $P(X = x) = 0 \quad \forall x$-re;
- $P(a < X \leq b) = P(a \leq X < b) = F(b) - F(a)$.

Abszolút folytonos val.változó várható értéke: $EX = \int\limits_{-\infty}^{\infty} xf(x)\,dx$.

Abszolút folytonos val.változó l. momentuma: $EX^l = \int\limits_{-\infty}^{\infty} x^l f(x)\,dx$.

Nevezetes abszolút folytonos eloszlások:

Eloszlás neve	Jelölése	Eloszlásfüggvény	Sűrűségfüggvény	EX	D^2X
Egyenletes	E(a,b)	$\begin{cases} 0 & \text{ha } x \leq a \\ \frac{x-a}{b-a} & \text{ha } a < x \leq b \\ 1 & \text{ha } b < x \end{cases}$	$\begin{cases} \frac{1}{b-a} & \text{ha } a < x \leq b \\ 0 & \text{különben} \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponenciális	Exp(λ)	$\begin{cases} 1-e^{-\lambda x} & \text{ha } x \geq 0 \\ 0 & \text{különben} \end{cases}$	$\begin{cases} \lambda e^{-\lambda x} & \text{ha } x \geq 0 \\ 0 & \text{különben} \end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Standard normális	N$(0,1^2)$	$\Phi(x) = ...$	$\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \quad x \in \mathbb{R}$	0	1
Normális	N(m,σ^2)	...	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}} \quad x \in \mathbb{R}$	m	σ^2

Állítás. **Val.változó függvényének várható értéke**

Legyen X val. változó; g: $\mathbb{R} \to \mathbb{R}$ függvény.
Ekkor

- $E(g(X)) = \sum_k g(x_k)p_k$, ha X diszkrét

- $E(g(X)) = \int\limits_{-\infty}^{\infty} g(x)f(x)\,dx$, ha X abszolút folytonos

Mindkét esetben a várható érték létezéséhez a szumma/integrál abszolút konvergenciájára van szükség.

Állítás. **Abszolút folytonos val.változó szigorúan monoton függvényének eloszlás- és sűrűségfüggvénye**
Legyen X abszolút folytonos val. változó; g: $\mathbb{R} \to \mathbb{R}$ szigorúan monoton, folytonosan differenciálható függvény. Ekkor
a.) Y=g(X) eloszlásfüggvénye:

$$F_Y(y) = F_{g(X)}(y) = \begin{cases} F_X(g^{-1}(y)) & \text{ha g szig. mon. növő} \\ 1 - F_X(g^{-1}(y)) & \text{ha g szig. mon. csökkenő} \end{cases}$$

b.) Y=g(X) sűrűségfüggvénye:

$$f_Y(y) = f_{g(X)}(y) = f_X(g^{-1}(y)) \cdot \left|[g^{-1}(y)]'\right| = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|}$$

Állítás. **Normálás**
Legyen X$\sim N(m, \sigma^2)$. Ekkor $\frac{X-m}{\sigma} \sim N(0,1)$.

Állítás. $\Phi(-x) = 1 - \Phi(x)$

Állítás. $\Phi^{-1}(q) = -\Phi^{-1}(1-q) \quad 0 < q < 1$

Definíció. **Val.változók konvolúciója:** Legyenek X és Y független val.változók. X és Y konvolúciójának (jel. X$*$Y) az X+Y val.változót nevezzük.
Állítás. **A konvolúció eloszlásának meghatározása**

- Diszkrét eset: $P(X+Y=k) = \sum\limits_{l=0}^{\infty} P(X=l) \cdot P(Y=k-l)$
- Folytonos eset: $f_{X+Y}(z) = \int\limits_{-\infty}^{\infty} f_X(u)f_Y(z-u)du$

Állítás. Legyenek $X_1, ..., X_n$, X és Y független val. változók
- $X_1 \sim \mathrm{Ind}(p), ..., X_n \sim \mathrm{Ind}(p) \quad \Rightarrow \quad X_1 + ... + X_n \sim \mathrm{Bin}(n,p)$
- $X \sim \mathrm{Bin}(n,p)$, $Y \sim \mathrm{Bin}(m,p) \quad \Rightarrow \quad X+Y \sim \mathrm{Bin}(n+m,p)$
- $X_1 \sim \mathrm{Geo}(p), ..., X_n \sim \mathrm{Geo}(p) \quad \Rightarrow \quad X_1 + ... + X_n \sim \mathrm{NegBin}(n,p)$
- $X \sim \mathrm{Poi}(\lambda_1)$, $Y \sim \mathrm{Poi}(\lambda_2) \quad \Rightarrow \quad X+Y \sim \mathrm{Poi}(\lambda_1+\lambda_2)$
- $X \sim N(m_1, \sigma_1^2)$, $Y \sim N(m_2, \sigma_2^2) \quad \Rightarrow \quad X+Y \sim N(m_1+m_2, \sigma_1^2+\sigma_2^2)$

Sűrűségfüggvény becslése magfüggvény segítségével n elemű mintából:
Parzen-Rosenblatt becslés: $f(x) \approx f_n(x) = \frac{1}{nh_n} \sum\limits_{i=1}^{n} k\left(\frac{x-X_i}{h_n}\right)$

Állítás. Az $F_n(x) = \sum\limits_{i=1}^{n} \frac{\mathrm{I}(X_i<x)}{n}$ tapasztalati eloszlásfüggvény 1 valószínűséggel tart a (valódi) F(x) eloszlásfüggvényhez.

Definíció. **Torzítatlan becslés:**
T($\mathbf{X}$) statisztika torzítatlan becslése θ-nak, ha $E_\theta T(\mathbf{X}) = \theta \quad \forall\theta$-ra.

Definíció. **Likelihood függvény:** Legyen $\mathbf{X} = (X_1, ..., X_n)$ i.i.d. minta
- $L(\theta, \mathbf{x}) = f_\theta(\mathbf{x}) = \prod\limits_{i=1}^{n} f_\theta(x_i)$, ha az eloszlás folytonos
- $L(\theta, \mathbf{x}) = P_\theta(\mathbf{X} = \mathbf{x}) = \prod\limits_{i=1}^{n} P_\theta(X_i = x_i)$, ha az eloszlás diszkrét.

Definíció. **Log-likelihood függvény:** $l(\theta, \mathbf{x}) = \log(L(\theta, \mathbf{x}))$.

Paraméterbecslési módszerek
- **Maximum likelihood módszer (ML-módszer):** Azt a paraméterértéket keressük, ahol a likelihood függvény a legnagyobb értéket veszi fel: $\max\limits_\theta L(\theta, \mathbf{x})$
  Amennyiben a függvény deriválható θ szerint, akkor a maximumot kereshetjük a szokásos módon, az első és második deriváltak segítségével, azonban a feladatunkat jelentősen megnehezíti, hogy olyan n-szeres szorzatot kellene deriválni, amelyiknek minden tagjában ott van az a változó, ami szerint deriválnunk kellene. Ezért likelihood függvény helyett a log-likelihood függvény maximumhelyét keressük.
  Ha θ 1 dimenziós, akkor az
  - elsőrendű feltétel: $\partial_\theta l(\theta, \mathbf{x}) = 0 \quad \rightsquigarrow \quad \hat\theta$
  - másodrendű feltétel: $\partial^2_\theta l(\theta, \mathbf{x}) < 0$

  Ha θ p dimenziós, akkor $\theta = (\theta_1, ..., \theta_p)$, az
  - elsőrendű feltétel:
    $\partial_{\theta_i} l(\theta, \mathbf{x}) = 0 \quad \rightsquigarrow \quad \hat\theta_i \ \ (i = 1, ..., p) \quad \rightsquigarrow \quad \hat\theta = (\hat\theta_1, ..., \hat\theta_p)$
  - másodrendű feltétel: $H(\theta_1, ..., \theta_p) = \left(\partial_{\theta_i}\partial_{\theta_j} l(\theta, \mathbf{x})\right)_{i,j=1,...,p}$ Hesse-mátrix negatív definit a $\theta = \hat\theta$ helyen

- **Momentum módszer:** A mintából számítható tapasztalati momentumokat ($m_i := \frac{\sum_j x_j^i}{n}$) egyenlővé tesszük az elméleti momentumokkal ($M_i := E_\theta X^i$), az elsőtől kezdve, mégpedig annyit, amennyi paraméter van. Tehát p darab ismeretlen paraméter esetén a következő p ismeretlenes egyenletrendszert oldjuk meg:
  $M_1 = m_1$
  $\vdots$
  $M_p = m_p$
  Megjegyzés: $m_1 = \overline{x}$

Fisher-tétel: Ha θ ML-becslése $\hat{\theta}$, akkor tetszőleges g függvény esetén $g(\theta)$ ML-becslése $g(\hat{\theta})$.

Definíció. **χ^2-eloszlás:** Az X valószínűségi változó n szabadságfokú χ^2-eloszlást követ (jel.: $X \sim \chi^2_n$), ha $X = U_1^2 + ... + U_n^2$, ahol $U_i \sim N(0,1)$ minden i-re és függetlenek egymástól.
Definíció. **t-eloszlás:** Az X valószínűségi változó n szabadságfokú Student-féle t-eloszlást követ (jel.: $X \sim t_n$), ha $X = \frac{Z}{\sqrt{\frac{Y_n}{n}}}$, ahol $Z \sim N(0,1)$ és $Y_n \sim \chi^2_n$ függetlenek egymástól.
Mostantól α egy 0-hoz közeli pozitív szám lesz (például $0.05 = 5\%$), és vezessük be a következő jelöléseket:

- u_α: $N(0,1)$ eloszlás $(1-\alpha)$-kvantilise, azaz $u_\alpha = \Phi^{-1}(1-\alpha)$
- $z_\alpha := u_{1-\alpha}$ (sok könyvben ezt használják)
- $t_{n,\alpha}$: n szabadságfokú t-eloszlás $(1-\alpha)$-kvantilise
- $\chi^2_{n,\alpha}$: n szabadságfokú χ^2-eloszlás α-kvantilise

Definíció. **Konfidencia intervallum:** Adott α-hoz legalább $(1-\alpha)$ valószínűséggel tartalmazza az adott paramétert (vagy annak egy függvényét): $P_\theta\left(T_1(\mathbf{X}) < \hat{\theta} < T_2(\mathbf{X})\right) \geq 1-\alpha$.
Gyakran keresünk szimmetrikus konfidencia intervallumot, ilyenkor $T_1 = T_2 =: \Delta$, és az intervallum $\hat{\theta} \pm \Delta$ alakba írható.

Legyen $X_1, ..., X_n \sim N(m, \sigma)$ i.i.d. minta

- m-re konfidencia intervallum
  - ha σ ismert, akkor $\overline{x} \pm u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
  - ha σ ismeretlen, akkor $\overline{x} \pm t_{n-1,\frac{\alpha}{2}} \frac{s_n^*}{\sqrt{n}}$
- σ^2-re konfidencia intervallum: $\left[\frac{(n-1)\cdot(s_n^*)^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}; \frac{(n-1)\cdot(s_n^*)^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right]$

Konfidencia intervallum a valószínűségre (p) nagy minta esetén, ha normális eloszlással közelítünk: $\hat{p} \pm u_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Definíció. **Valószínűségi vektorváltozó:** $\mathbf{X}: \Omega \to \mathbb{R}^d$ mérhető függvény, azaz amire $\{\omega : \mathbf{X}(\omega) \in B\} \in \mathcal{A}$ minden B$\subseteq \mathbb{R}^d$ nyílt halmazra.

Definíció. **Valószínűségi vektorváltozó eloszlása:**
$Q_{\mathbf{X}}(B) = P(\mathbf{X} \in B) = P(\omega : \mathbf{X}(\omega) \in B)$

Definíció. **X valószínűségi vektorváltozó eloszlásfüggvénye:**
$F_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} < \mathbf{x}) = P(X_1 < x_1, ..., X_d < x_d)$.

Állítás. **Az eloszlásfüggvény tulajdonságai:**

- $0 \leq F_{\mathbf{X}}(\mathbf{x}) \leq 1$;
- minden koordinátájában monoton növő;
- minden koordinátájában balról folytonos;
- $\lim\limits_{x_1,...,x_d \to \infty} F_{\mathbf{X}}(x_1, ..., x_d) = 1$;
- $\lim\limits_{x_i \to -\infty} F_{\mathbf{X}}(x_1, ..., x_d) = 0$ minden i-re.

Definíció. **X valószínűségi vektorváltozó abszolút folytonos** , ha létezik olyan $f_{\mathbf{X}}(x_1, ..., x_d)$ függvény, amelyre
$F_{\mathbf{X}}(x_1, ..., x_d) = \int\limits_{-\infty}^{x_1} ... \int\limits_{-\infty}^{x_d} f_{\mathbf{X}}(t_1, ..., t_d)\, dt_1 ... dt_d$.
Ilyenkor $f_{\mathbf{X}}(\mathbf{x})$-et **sűrűségfüggvény**nek hívjuk.

Mostantól $d = 2$ lesz, és a következő jelöléseket és elnevezéseket használjuk:

- $F_{X,Y}(x,y) = P(X < x, Y < y)$ $\rightsquigarrow$ együttes eloszlásfüggvény
- $F_X(x) = P(X < x)$, $F_Y(y) = P(Y < y)$ $\rightsquigarrow$ peremeloszlásfüggvények
- $f_{X,Y}(x,y)$ $\rightsquigarrow$ együttes sűrűségfüggvény
- $f_X(x), f_Y(y)$ $\rightsquigarrow$ peremsűrűségfüggvények

Állítás.

- $F_X(x) = \lim\limits_{y \to \infty} F_{X,Y}(x,y)$ és $F_Y(y) = \lim\limits_{x \to \infty} F_{X,Y}(x,y)$

- $F_{X,Y}(x,y) = \int\limits_{-\infty}^{y} \int\limits_{-\infty}^{x} f_{X,Y}(u,v)\, dudv$
- $f_{X,Y}(x,y) = \partial_y \partial_x F_{X,Y}(x,y) = \partial_x \partial_y F_{X,Y}(x,y)$
- $\int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} f_{X,Y}(x,y)\, dxdy = 1$
- $f_X(x) = \int\limits_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy$ és $f_Y(y) = \int\limits_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx$

Állítás.

- X,Y függetlenek $\Leftrightarrow$ $F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y)$
- X,Y függetlenek $\Leftrightarrow$ $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$
- X,Y függetlenek $\Leftrightarrow$ $P(X=x, Y=y) = P(X=x) \cdot P(Y=y)$
- X,Y függetlenek $\Rightarrow$ $E(XY) = EX \cdot EY$

Definíció. X **és** Y **kovarianciája:** $\mathrm{Cov}(X,Y) = E\left[(X-EX)(Y-EY)\right]$.
Köv.: $\mathrm{Cov}(X,Y) = E(XY) - EXEY$.
Elnevezés: ha $\mathrm{Cov}(X,Y) = 0$, akkor azt mondjuk, hogy X és Y **korrelálatlanok**.

Állítás.

- Ha X és Y függetlenek egymástól, akkor korrelálatlanok is.
- Ha X és Y korrelálatlanok, akkor ebből **nem** következik, hogy függetlenek is!!!!!

Állítás. **A kovariancia tulajdonságai:**
Legyenek $X, Y, X_1, ..., X_n$ valószínűségi változók, $a, b \in \mathbb{R}$. Ekkor

- $\mathrm{Cov}(X,X) = D^2X$
- $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
- $\mathrm{Cov}(X,a) = 0$
- $\mathrm{Cov}(aX,bY) = ab\mathrm{Cov}(X,Y)$
- $D^2(X+Y) = D^2X + D^2Y + 2\mathrm{Cov}(X,Y)$
- $D^2(\sum\limits_{i=1}^{n} X_i) = \sum\limits_{i=1}^{n} D^2X_i + 2 \sum\limits_{1\le i<j\le n} \mathrm{Cov}(X_i,X_j)$
- X,Y függetlenek $\Rightarrow$ $\mathrm{Cov}(X,Y) = 0$

Definíció. X **és** Y **korrelációja:** $R(X,Y) = \frac{\mathrm{Cov}(X,Y)}{DXDY}$.
A korreláció két valószínűségi változó lineáris kapcsolatát méri:

- $R > 0$ $\Rightarrow$ pozitív a kapcsolat
  $R < 0$ $\Rightarrow$ negatív a kapcsolat
- $R^2 \sim 1$ $\Rightarrow$ erős a kapcsolat
  $R^2 \sim 0.5$ $\Rightarrow$ közepes a kapcsolat
  $R^2 \sim 0$ $\Rightarrow$ gyenge a kapcsolat

Legyenek X és Y valószínűségi változók. Y-nak X-re vonatkozó feltételes várható értéke – $\mathrm{E}(Y|X)$ – precíz definiálására nem vállalkozok, úgy gondoljunk rá, mint egy valószínűségi változóra; és ha X egy adott értéket vesz fel – $\mathrm{E}(Y|X=x)$ –, akkor mint konkrét számra.
$\mathrm{E}(Y|X)$ abszolút folytonos eloszlások esetén a következő képlettel számítható: $\mathrm{E}(Y|X) = \left.\int\limits_{-\infty}^{\infty} y f_{Y|X}(y|x) dy\right|_{x=X}$
ahol $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$ a feltételes sűrűségfüggvény.
Nadarajah-módszer a feltételes várható érték becslésére $(X_1,Y_1), ..., (X_n,Y_n)$ minta alapján: $E(Y|X=x) \approx \frac{\sum_i Y_i k\left(\frac{x-X_i}{h_n}\right)}{\sum_i k\left(\frac{x-X_i}{h_n}\right)}$.

Állítás. Legyen g mérhető függvény.

- $E[g(X)|X] = g(X)$
- X,Y függetlenek $\Rightarrow E(Y|X) = EY$

<u>Feladat</u>: Y val. változót szeretnénk közelíteni X val. változó tetszőleges függvénye segítségével:

$$\mathrm{E}[Y - f(X)]^2 \longrightarrow \min_f \qquad \rightsquigarrow \text{Megoldása: } f_{opt} = \mathrm{E}(Y|X)$$

<u>Feladat</u>: Y val. változót szeretnénk közelíteni X val. változó lineáris függvénye segítségével:

$$\mathrm{E}[Y - (aX+b)]^2 \longrightarrow \min_{a,b} \qquad \rightsquigarrow \text{Megoldása: } \begin{aligned} a_{opt} &= \frac{Cov(X,Y)}{D^2(X)} \\ b_{opt} &= EY - a_{opt}EX \end{aligned}$$

<u>Feladat</u> (**lineáris modell**): Adottak $(x_1,y_1), \ldots, (x_n,y_n)$ pontok, ezekre szeretnénk egyenest illeszteni (neve: *regressziós egyenes*) legkisebb négyzetek módszerével.
A modell: $Y_i = aX_i + b + \varepsilon_i$, ahol $\mathrm{E}\varepsilon_i = 0$ és $\mathrm{D}^2\varepsilon_i = \sigma^2 < \infty$ $(i = 1, \ldots, n)$

Megoldás: $\hat{a} = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2}$, $\hat{b} = \overline{y} - \hat{a}\overline{x}$
Reziduumok: $\hat{\varepsilon}_i = y_i - \hat{a}x_i - \hat{b}$ (i=1,..., n)
Reziduális négyzetösszeg: RNÖ=$\sum \hat{\varepsilon}_i^2 = \sum(y_i - \overline{y})^2 - \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2}$
$\hat{\sigma}^2 = \frac{\text{RNÖ}}{n-2}$

Tétel. **Markov-egyenlőtlenség:** Legyen $g : \mathbb{R} \to \mathbb{R}$ monoton növő függvény, $X \geq 0$ val.változó, $\varepsilon > 0$ tetsz.
Ekkor $\mathrm{P}(X \geq \varepsilon) \leq \frac{E[g(X)]}{g(\varepsilon)}$.
Spec., ha $g(x) = x \quad \Rightarrow \quad \mathrm{P}(X \geq \varepsilon) \leq \frac{E(X)}{\varepsilon}$
Tétel. **Csebiseb-egyenlőtlenség:** $\mathrm{P}(|X - EX| \geq \varepsilon) \leq \frac{D^2(X)}{\varepsilon^2}$.

Legyen $X_1, X_2, \ldots$ valószínűségi változók sorozata.
Definíció. **1 valószínűségű konvergencia:**
$X_n \overset{n\to\infty}{\longrightarrow} X$ *1 valószínűséggel*, ha $P(\{\omega : X_n(\omega) \overset{n\to\infty}{\to} X(\omega)\}) = 1$.
Definíció. **Gyenge konvergencia:** $X_n \overset{n\to\infty}{\longrightarrow} X$ *gyengén*, ha az eloszlásfüggvényeikre $F_n(x) \overset{n\to\infty}{\longrightarrow} F(x)$ $\quad F$ minden folytonossági pontjában.

Tétel. **Nagy számok törvénye (NSZT):**
Legyenek $X_1, X_2, \ldots$ i.i.d. val. változók, $\mathrm{E}X_1 = m < \infty$.
Ekkor $\frac{X_1+\ldots+X_n}{n} \overset{n\to\infty}{\to} m$ 1 valószínűséggel.

Tétel. **Centrális határeloszlás tétel (CHT):**
Legyenek $X_1, X_2, \ldots$ i.i.d. val. változók, $\mathrm{E}X_1 = m$, $D^2(X_1) = \sigma^2 < \infty$. Ekkor
$\frac{X_1+\ldots+X_n-nm}{\sqrt{n}\sigma} \overset{n\to\infty}{\to}$ N(0,1) gyengén, azaz $P\left(\frac{X_1+\ldots+X_n-nm}{\sqrt{n}\sigma} < x\right) \overset{n\to\infty}{\to} \Phi(x)$.

Hipotézis $\sim$ valami állítás, aminek igazságát vizsgálni szeretnénk

Paramétertér: $\Theta = \Theta_0 \cup^* \Theta_1 \longrightarrow$ "valóság"
Mintatér: $\mathcal{X} = \mathcal{X}_e \cup^* \mathcal{X}_k \longrightarrow$ "látszat" - MINTÁBÓL
$\mathcal{X}_k$: kritikus tartomány - azon $\mathbf{X}$ megfigyelések halmaza, amikre *elutasítjuk* a nullhipotézist
$\mathcal{X}_e$: elfogadási tartomány - azon $\mathbf{X}$ megfigyelések halmaza, amikre *elfogadjuk* a nullhipotézist

Hipotézisvizsgálati feladat:
$H_0 : \theta \in \Theta_0 \quad \rightsquigarrow$ nullhipotézis
$\underline{H_1 : \theta \in \Theta_1} \quad \rightsquigarrow$ ellenhipotézis

Tehát ha $\mathbf{X} \in \mathcal{X}_e$, akkor elfogadjuk H_0-t; ha $\mathbf{X} \in \mathcal{X}_k$, akkor pedig elutasítjuk H_0-t.
Amennyiben a Θ_0 halmaz egyelemű, akkor azt mondjuk, hogy H_0 egyszerű. H_1-re ugyanígy.

Az $\mathcal{X}$ mintatér felosztását általában egy statisztika (neve: próbastatisztika) segítségével végezzük el:
legyen T: $\mathcal{X} \to \mathbb{R}$, $\mathcal{X}_k = \{\underline{x} \in \mathcal{X} : \mathrm{T}(\underline{x}) > c\}$ $\quad c$ neve: kritikus érték
$\mathcal{X}_e = \{\underline{x} \in \mathcal{X} : \mathrm{T}(\underline{x}) \leq c\}$

döntés / "valóság"	H_0-t elfogadjuk ($\mathcal{X}_e$)	H_0-t elutasítjuk ($\mathcal{X}_k$)
H_0 teljesül (Θ_0)	helyes döntés	elsőfajú hiba
H_0 nem teljesül (Θ_1)	másodfajú hiba	helyes döntés

P(elsőfajú hiba)=$\alpha(\theta)$=$\mathrm{P}_\theta(\mathcal{X}_k)$, ahol $\theta \in \Theta_0$
P(másodfajú hiba)=$\beta(\theta)$=$\mathrm{P}_\theta(\mathcal{X}_e)$, ahol $\theta \in \Theta_1$
Erőfüggvény: ψ: $\Theta_1 \to \mathbb{R}$, $\psi(\theta) = P_\theta(\mathcal{X}_k)$
Terjedelem: α = sup $\{\alpha(\theta)$: $\theta \in \Theta_0\}$

Azt mondjuk, hogy az 1-es próba *erősebb* a 2-es próbánál, ha $\alpha_1 = \alpha_2$ és $\psi_1(\theta) \geq \psi_2(\theta)\ \forall\theta \in \Theta_1$.
Próbafüggvény: φ: $\mathcal{X} \to$[0,1] $\rightsquigarrow$ ennyi valószínűséggel vetem el a H_0-t a minta alapján
$\mathbf{x} \in \mathcal{X}_k \Rightarrow \varphi(\underline{x}) = 1$
$\mathbf{x} \in \mathcal{X}_e \Rightarrow \varphi(\underline{x}) = 0$

p-érték: az az α terjedelem, ami esetén a próbastatisztika értéke egyenlő a kritikus értékkel : T($\mathbf{x}$)= c_α.
A p-érték a legkisebb terjedelem, amire még elutasítjuk a H_0-t. Ha egy próbát számítógép segítségével végzünk el, rendszerint a p-érték révén tudunk dönteni: ha (p-érték)< α, akkor elvetjük H_0-t.

Ha mind H_0, mind H_1 egyszerű, akkor adott α terjedelemhez lehet legerősebb próbát találni, ezt pedig úgy hívják, hogy *valószínűség-hányados próba.* A hipotéziseket folytonos esetre írom fel. Diszkrétre a sűrűségfüggvény helyett a konkrét eloszlást kell írni.

$H_0: \; f = f_0$
$H_1: \; f = f_1$

A valószínűség-hányados próba kritikus tartománya: $\mathcal{X}_k = \left\{ \mathbf{x} : \frac{f_1(x)}{f_0(x)} > c_\alpha \right\}$

Tehát azokat az $\mathbf{x}$-eket, amire az $\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})}$ nagy, bepakoljuk a kritikus tartományba egészen addig, míg az adott α terjedelmet el nem érjük. Diszkrét esetben ehhez általában véletlenítésre van szükség, azaz bizonyos $\mathbf{x}$-ek esetén nem 1 vagy 0, hanem egy, e két szám közé eső (jelöljük p_α-val) valószínűséggel vetjük el a nullhipotézist.

Néhány konkrét próba – az α végig a próba terjedelmét jelöli, ami előre adott

1.) Egymintás próbák

a.) Egymintás u-próba

$X_1, \dots, X_n \sim$N(m, σ^2), ahol σ ismert, m paraméter

a.) $H_0: m = m_0$, $H_1: m \neq m_0$ b.) $H_0: m = m_0$, $H_1: m > m_0$ c.) $H_0: m = m_0$, $H_1: m < m_0$

A próbastatisztika: $\mathrm{T}(\mathbf{X}) = u = \sqrt{n}\frac{\overline{X} - m_0}{\sigma} \overset{H_0 \text{ esetén}}{\sim} N(0,1)$

A kritikus tartományok:

a.) $\mathcal{X}_k = \{\mathbf{x} : |u| > u_{\alpha/2}\}$
b.) $\mathcal{X}_k = \{\mathbf{x} : u > u_\alpha\}$
c.) $\mathcal{X}_k = \{\mathbf{x} : u < -u_\alpha\}$

b.) Egymintás t-próba

$X_1, \dots, X_n \sim$N(m, σ^2), ahol σ, m paraméter

a.) $H_0: m = m_0$, $H_1: m \neq m_0$ b.) $H_0: m = m_0$, $H_1: m > m_0$ c.) $H_0: m = m_0$, $H_1: m < m_0$

A próbastatisztika: $\mathrm{T}(\mathbf{X}) = t = \sqrt{n}\frac{\overline{X} - m_0}{s_n^*} \overset{H_0 \text{ esetén}}{\sim} \mathrm{t}_{n-1}$

A kritikus tartományok:

a.) $\mathcal{X}_k = \{\mathbf{x} : |t| > t_{n-1,\alpha/2}\}$
b.) $\mathcal{X}_k = \{\mathbf{x} : t > t_{n-1,\alpha}\}$
c.) $\mathcal{X}_k = \{\mathbf{x} : t < -t_{n-1,\alpha}\}$

2.) Kétmintás próbák

$X_1, \dots, X_n \sim$N(m_1, σ_1^2)
$Y_1, \dots, Y_m \sim$N(m_2, σ_2^2)

Az elvégzendő próbák $H_0: m_1 = m_2$ nullhipotézis esetén:

	a két minta független		a két minta nem független
σ_1 és σ_2 ismert	b.) kétmintás u-próba		egymintás u-próba a különbségekre
σ_1 és σ_2 ismeretlen	előzetes F-próba		egymintás t-próba a különbségekre
	$\sigma_1 = \sigma_2$	$\sigma_1 \neq \sigma_2$	
	c.) kétmintás t-próba	d.) Welch-próba	

a.) F-próba

$m_1, m_2, \sigma_1, \sigma_2$ paraméterek

$H_0: \sigma_1 = \sigma_2$ és H_1: ami a szövegkörnyezetben értelmes

A próbastatisztika: $F = \begin{cases} \frac{(s_1^*)^2}{(s_2^*)^2} \overset{H_0 \text{ esetén}}{\sim} F_{n-1,m-1} & \text{ha } s_1^* > s_2^* \\ \frac{(s_2^*)^2}{(s_1^*)^2} \overset{H_0 \text{ esetén}}{\sim} F_{m-1,n-1} & \text{ha } s_2^* > s_1^* \end{cases}$

b.) kétmintás u-próba

m_1, m_2 paraméterek, σ_1, σ_2 ismert

$H_0: m_1 = m_2$ és H_1: ami a szövegkörnyezetben értelmes

A próbastatisztika: $u = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \overset{H_0 \text{ esetén}}{\sim}$N(0,1)

c.) kétmintás t-próba

$m_1, m_2, \sigma_1 = \sigma_2$ paraméterek

$H_0: m_1 = m_2$ és H_1: ami a szövegkörnyezetben értelmes

A próbastatisztika: $t = \sqrt{\frac{nm}{n+m}} \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{(n-1)(s_1^*)^2 + (m-1)(s_2^*)^2}{n+m-1}}} \overset{H_0 \text{ esetén}}{\sim} \mathrm{t}_{n+m-2}$

d.) Welch-próba

$m_1, m_2, \sigma_1 \neq \sigma_2$ paraméterek

$H_0: m_1 = m_2$ és H_1: ami a szövegkörnyezetben értelmes

A próbastatisztika: $t' = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{(s_1^*)^2}{n} + \frac{(s_2^*)^2}{m}}} \overset{H_0 \text{ esetén}}{\sim} \mathrm{t}_f$, ahol

$\frac{1}{f} = \frac{c^2}{n-1} + \frac{(1-c)^2}{m-1}$

$c = \frac{\frac{(s_1^*)^2}{n}}{\frac{(s_1^*)^2}{n} + \frac{(s_2^*)^2}{m}}$, ha $s_1^* > s_2^*$

3.) χ^2-próbák

a.) Diszkrét illeszkedésvizsgálat

Feladat: adott egy $\mathbf{X} = (X_1, \dots, X_n)$ n elemű minta, és azt akarjuk eldönteni, hogy a minta egy általunk "remélt" eloszlásból származik-e. *Diszkrét* illeszkedésvizsgálatnál feltesszük, hogy a mintaelemek r különböző értéket vehetnek fel: $\mathrm{P}(X_i = x_j) = p_j \quad j = 1, \ldots, r$. Jelöljük N_j-vel a gyakoriságokat, azaz azt, hogy az n elemű mintában hány darab x_j szerepel.

Osztályok	1	2	...	r	Összesen
Valószínűségek	p_1	p_2	...	p_r	1
Gyakoriságok	N_1	N_2	...	N_r	n

H_0 : a valószínűségek: $\mathbf{p}=(p_1, \ldots, p_r)$

H_1: nem ezek a valószínűségek

A próbastatisztika: $T_n = \sum\limits_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i} \xrightarrow{H_0 \text{ esetén}} \chi^2_{r-1}$ eloszlásban, ha $n \to \infty$

A kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : T_n(\mathbf{x}) > \chi^2_{r-1,1-\alpha}\}$

Becsléses illeszkedésvizsgálat: csak annyit "sejtünk", hogy a minta valamilyen eloszlású, viszont a paramétereiről nincs sejtésünk. Ilyenkor amennyiben ML-módszerrel becsüljük meg az s darab ismeretlen paramétert, akkor a próbastatisztika: $T_n \xrightarrow{H_0 \text{ esetén}} \chi^2_{r-1-s}$ eloszlásban, ha $n \to \infty$.

b.) Függetlenségvizsgálat

Feladat: van egy minta, két szempont szerint csoportosítva. Azt kell eldönteni, hogy a két szempont független-e egymástól.

$\mathrm{p}_{i,j}$=P(egy megfigyelés az (i,j) osztályba kerül)

$\mathrm{N}_{i,j}$=ennyi megfigyelés kerül az (i,j) osztályba

A mintavétel eredménye:

		2. szempont					
		1	...	j	...	s	Összesen
	1	N_{11}	...	N_{1j}	...	N_{1s}	$N_{1\bullet}$
	⋮	⋮		⋮		⋮	⋮
1. szempont	i	N_{i1}	...	N_{ij}	...	N_{is}	$N_{i\bullet}$
	⋮	⋮		⋮		⋮	⋮
	r	N_{r1}	...	N_{rj}	...	N_{rs}	$N_{r\bullet}$
Összesen		$N_{\bullet 1}$	...	$N_{\bullet j}$	...	$N_{\bullet s}$	n

$$N_{i\bullet} = \sum_{j=1}^{s} N_{i,j}$$

$$N_{\bullet j} = \sum_{i=1}^{r} N_{i,j}$$

H_0 : a szempontok függetlenek, azaz $p_{i,j} = p_{i\bullet} \cdot p_{\bullet j} \ \forall i, j$-re

H_1: nem azok

A próbastatisztika: $T_n = n \left(\sum\limits_{i=1}^{r} \sum\limits_{j=1}^{s} \frac{N_{i,j}^2}{N_{i\bullet} N_{\bullet j}} - 1 \right) \xrightarrow{H_0 \text{ esetén}} \chi^2_{(r-1)(s-1)}$ eloszlásban, ha $n \to \infty$

A kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : T_n(\underline{x}) > \chi^2_{(r-1)(s-1),1-\alpha}\}$