Segédanyag a Valószínűségszámítás és statisztika tantárgyhoz

2015. szeptember 16.

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?
$P(A|B) = \frac{P(A\cap B)}{P(B)}$ $(P(B) \neq 0)$

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. **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:** $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$.

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

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
- Poisson: ritka események bekövetkezését írja le

Legyen X diszkrét valószínűségi változó, ami az x_1,x_2,... értékeket veszi fel, $p_1, p_2, \ldots$ valószínűségekkel.

Definíció. X **várható értéke** : $EX = \sum_{i=1}^{\infty} x_k p_k$, ha a végtelen összeg abszolút konvergens.

Definíció. X **l. momentuma** : $EX^l = \sum_{i=1}^{\infty} (x_k)^l p_k$, ha a végtelen összeg abszolút konvergens.

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

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

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:

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

Állítás. Tetszőleges X val.változó esetén

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

Nevezetes abszolút folytonos eloszlások:

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

Állítás. Legyen X abszolút folytonos eloszlású. Ekkor

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

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

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

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

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

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

Állítás. $\Phi(-x) = 1 - \Phi(x)$

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

2 dimenziós valószínűségi vektorváltozók:

- $F_{X,Y}(x,y) = P(X < x, Y < y)$ $\rightsquigarrow$ 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.

- $F_X(x) = \lim_{y\to\infty} F_{X,Y}(x,y)$ és $F_Y(y) = \lim_{x\to\infty} F_{X,Y}(x,y)$
- $F_{X,Y}(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u,v)\,du\,dv$
- $f_{X,Y}(x,y) = \partial_y\partial_x F_{X,Y}(x,y) = \partial_x\partial_y F_{X,Y}(x,y)$
- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1$
- $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$ és $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx$

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

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

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

Minta: $X_1, ..., X_n$ valószínűségi változó sorozat. Jel. $\mathbf{X} = (X_1, ..., X_n)$
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

Néhány lényeges statisztika:
- *Rendezett minta*: $X_1^* \le ... \le X_n^*$ nem csökkenő sorrendbe tesszük a mintaelemeket
- *Mintaátlag*: $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
- *Tapasztalati szórás*: $S_n = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n}}$
  Értelmezése: az átlagtól való átlagos eltérés abszolút mértékegységben
- *Korrigált tapasztalati szórás*: $S_n^* = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}}$
- *Szórási együttható (relatív szórás)*: $V = \frac{S_n}{\overline{X}}$
  Értelmezése: az átlagtól való átlagos eltérés százalékban
- *Tapasztalati eloszlásfüggvény*: $F_n(x) = \frac{\sum_{i=1}^{n} I(X_i<x)}{n}$
  ahol $I(X_i < x) = \begin{cases} 1 & \text{ha } X_i < x \\ 0 & \text{ha } X_i \ge x \end{cases} \rightsquigarrow$ karakterisztikus függvény
- *z-kvantilis*: $q_z = \inf\{x : F(x) \ge z\}$, 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

Definíció. **Torzítatlan becslés:**
$\mathrm{T}(\mathbf{X})$ 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

- $L(\theta, \mathbf{x}) = f_\theta(\mathbf{x}) = \prod_{i=1}^{n} f_\theta(x_i)$, ha az eloszlás folytonos
- $L(\theta, \mathbf{x}) = P_\theta(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{n} P_\theta(X_i = x_i)$, 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: $\max_\theta L(\theta, \mathbf{x})$
- **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.

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: $\left[\frac{(n-1)\cdot(s^*_n)^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}; \frac{(n-1)\cdot(s^*_n)^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right]$

p-érték: az az α terjedelem, ami esetén a próbastatisztika értéke egyenlő a kritikus értékkel : $\mathrm{T}(\mathbf{x}) = c_\alpha$.
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)$< \alpha$, akkor elvetjük H_0-t.

Néhány konkrét próba – az α végig a próba terjedelmét jelöli.

1.) Egymintás próbák

a.) Egymintás u-próba

$X_1, \ldots, X_n \sim \mathrm{N}(m, \sigma^2)$, ahol σ ismert, m paraméter

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

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

$X_1, \ldots, X_n \sim \mathrm{N}(m, \sigma^2)$, ahol σ, m paraméter

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

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.) $\mathcal{X}_k = \{\mathbf{x} : t < -t_{n-1,\alpha}\}$

3.) χ^2-próbák

a.) Diszkrét illeszkedésvizsgálat

H_0 : a valószínűségek: $\mathbf{p}=(p_1, \ldots, p_r)$
H_1: nem ezek a valószínűségek

A próbastatisztika: $T_n = \sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i} \overset{H_0 \text{ esetén}}{\longrightarrow} \chi^2_{r-1}$ 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: $T_n \overset{H_0 \text{ esetén}}{\longrightarrow} \chi^2_{r-1-s}$ eloszlásban, ha $n \to \infty$.

b.) Függetlenségvizsgálat

H_0 : a szempontok függetlenek
H_1: nem azok

Próbastatisztika: $T_n = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{N^2_{i,j}}{N_{i\bullet}N_{\bullet j}} - 1\right) \overset{H_0 \text{ esetén}}{\longrightarrow} \chi^2_{(r-1)(s-1)}$ eloszlásban $(n \to \infty)$
A kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : T_n(\underline{x}) > \chi^2_{(r-1)(s-1),1-\alpha}\}$