Záróvizsga tételsor

5. Valószínűségszámítási és statisztikai alapok

Fekete Dóra

Valószínűségszámítási és statisztikai alapok
Diszkrét és folytonos valószínűségi változók, nagy számok törvénye, centrális határeloszlás tétel. Statisztikai becslések, klasszikus statisztikai próbák.

1 Kolmogorov-féle valószínűségi mező

$(\Omega, \mathcal{A}, P)$ hármas, ahol:

Ω nem üres halmaz, eseménytér, ω elemi esemény.

$\mathcal{A}$ Ω részhalmazainak egy rendszere, $\mathcal{A} \subset 2^\Omega, A \in \mathcal{A}$ események, $\mathcal{A}$ σ-algebra.
σ-algebra:

1. $\Omega \in \mathcal{A}$
2. $A \in \mathcal{A} \Rightarrow \overline{A} = \Omega \backslash A \in \mathcal{A}$
3. $A_1, A_2 ... \in \mathcal{A} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$

P : $\mathcal{A} \to [0,1]$ halmazfüggvény, valószínűség, amelyre $P(\Omega) = 1, P(A) \geq 0, \forall A \in \mathcal{A}$-ra, páronként kizáró ($A_i \cdot A_j = \emptyset, i \neq j$) $A_1, A_2 ... \in \mathcal{A}$ eseményekre $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

2 Diszkrét és folytonos valószínűségi változók

- *Valószínűségi változó*: $\xi : \Omega \to \mathbb{R}$ mérhető függvény, azaz amire $\{\omega : \xi(\omega) < x\} \in \mathcal{A}, \forall x \in \mathbb{R}$, ahol $\mathcal{A}$ az eseménytér (Ω) részhalmazainak egy rendszere. ($\omega \in \Omega$ elemi esemény).
- *Valószínűségi változó eloszlása/eloszlásfüggvénye*: $F_\xi(x) = P(\xi < x)$, $\forall x \in \mathbb{R}$.
  Tulajdonságai:

  1. $0 \leq F_\xi(x) \leq 1$
  2. monoton növő
  3. balról folytonos
  4. $\lim\limits_{x \to -\infty} F_\xi(x) = 0$, $\lim\limits_{x \to \infty} F_\xi(x) = 1$

2.1 Diszkrét valószínűségi változók

Értékkészlete legfeljebb megszámlálhatóan végtelen, azaz $\{x_1 ... x_n ...\}$ elemekből áll. Ekkor eloszlása: $p_k := P(\xi = x_k)$.

Név	Értelmezés	Eloszlás	EX	D^2X
indikátor $Ind(p)$	Egy p valószínűségű esemény bekövetkezik-e vagy sem.	$P(X=1)=p$ $P(X=0)=1-p$	p	$p(1-p)$
geometriai (Pascal) $Geo(p)$	Hányadikra következik be először egy p valószínűségű esemény.	$P(X=k)=p(1-p)^{k-1}$ $k=1,2...$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
hipergeometriai $Hipgeo(N,M,n)$	Visszatevés nélküli mintavétel.	$P(X=k)=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$ $k=0,1,...,n$	$n\frac{M}{N}$	$n\frac{M}{N}(1-\frac{M}{N})(1-\frac{n-1}{N-1})$
binomiális $Bin(n,p)$	Visszatevéses mintavétel.	$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ $k=0,1,...,n$	np	$np(1-p)$
negatív binomiális $Negbin(n,p)$	Hányadikra következik be n. alkalommal egy p valószínűségű esemény.	$P(X=k)=\binom{k-1}{n-1}p^n(1-p)^{k-n}$ $k=n,n+1,...$	$\frac{n}{p}$	$\frac{n(1-p)}{p^2}$
Poisson $Poi(\lambda)$	Ritka esemény.	$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$	λ	λ

2.2 Folytonos valószínűségi változók

Egy ξ valószínűségi változó abszolút folytonos, ha létezik olyan $f(x)$ függvény, amelyre $F(x)=\int_{-\infty}^{x}f(t)dt$. Ilyenkor $f(x)$ *sűrűségfüggvény*. ($F(x)$ pedig az eloszlásfüggvény.)
Másik megfogalmazás: $\forall a<b$-re $P(a<\xi<b)=\int_a^b f(t)dt$, $F_\xi(x)=P(\xi<x)=\lim\limits_{a\to-\infty}P(a<\xi\ b)=\int_{-\infty}^{x}f(t)dt$.
Sűrűségfüggvény tulajdonságai:

1. $f(x)=F'(x)$
2. $f(x)\geq 0$
3. $\int_{-\infty}^{\infty}f(x)dx=1$

Név	Eloszlásfüggvény	Sűrűségfüggvény	EX	D^2X
egyenletes $E(a,b)$	$\begin{cases}0 & x\leq a\\ \frac{x-a}{b-a} & a<x\leq b\\ 1 & b<x\end{cases}$	$\begin{cases}\frac{1}{b-a} & a<x\leq b\\ 0 & otherwise\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
exponenciális $Exp(\lambda)$	$\begin{cases}1-e^{-\lambda x} & x\geq 0\\ 0 & otherwise\end{cases}$	$\begin{cases}\lambda\cdot e^{-\lambda x} & x\geq 0\\ 0 & otherwise\end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
normális $N(m,\sigma^2)$	...	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}\ x\in\mathbb{R}$	m	σ^2
standard normális $N(0,1^2)$	$\Phi(x)=...$	$\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\ x\in\mathbb{R}$	0	1
gamma $\Gamma(\alpha,\lambda)$	...	$\begin{cases}\frac{1}{\Gamma(\alpha)}\lambda^\alpha x^{\alpha-1}e^{-\lambda x} & x\geq 0\\ 0 & otherwise\end{cases}$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$

2.3 Fogalmak

- *Konvolúció*: X,Y független valószínűségi változók, konvolúciójuk az $X+Y$ v. v.
- *Függetlenség*: $P(\xi_1<x_1,...,\xi_n<x_n)=\prod_{i=1}^{n}P(\xi_i<x_i)$ vagy diszkrét esetben: $P(\xi_1)x_1,...,\xi_n=x_n)=\prod_{i=1}^{n}P(\xi_i=x_i)$
- *Várható érték*: $(\Omega,\mathcal{A},P)$ valószínűségi mező, $X:\Omega\to\mathbb{R}$ valószínűségi változó, $EX=\int_\Omega XdP$, ha ez létezik. Diszkrét esetben $EX=\sum_k x_k\cdot p_k$, ha abszolút konvergens. Abszolút folytonos esetben $EX=\int_{-\infty}^{\infty}x\cdot f(x)dx$, ha abszolút folytonos.
- *Szórásnégyzet*: $D^2X=E((X-EX)^2)=EX^2-E^2X$

- *l. momentum*: $EX^l = \int_\Omega x^l dP$, ha létezik.
- *Szórás*: $DX = \sqrt{D^2X}$
- *Kovariancia*: $cov(X,Y) = E((X-EX)(Y-EY) = E(XY) - EX \cdot EY$. Ha $cov(X,Y) = 0$, akkor X és Y korrelálatlan. (Megjegyzés: ha két v.v. független, akkor $cov(X,Y) = 0$, vagyis korrelálatlanok; illetve $cov(X,X) = D^2X$.)
- *Korreláció*: $R(X,Y) = \frac{cov(X,Y)}{DX \cdot DY}$, két v.v. lineáris kapcsolatát méri. $R > 0 \to$ pozitív, $R < 0 \to$ negatív; $R^2 \sim 1 \to$ erős, $R^2 \sim 0.5 \to$ közepes, $R^2 \sim 0 \to$ gyenge.

3 Nagy számok törvénye

3.1 Gyenge törvény

$X_1, X_2, ...$ függetlenek, azonos eloszlásúak, $EX_i = m < \infty$, $D^2X_i = \sigma^2 < \infty$.
$P(\frac{X_1+...+X_n}{n} - m \geq \varepsilon) \to 0$ $(n \to \infty)$ $\forall \varepsilon > 0$-ra (sztochasztikus konvergencia).

3.2 Erős törvény

$X_1, X_2, ...$ függetlenek, azonos eloszlásúak, $EX_1 = m < \infty$, $D^2X_1 = \sigma^2 < \infty$.
$\frac{X_1+...+X_n}{n} \to m$ $(n \to \infty)$ 1 valószínűséggel.
Megjegyzés: Csebisev-egyenlőtlenséggel bizonyítjuk. ($\frac{\sigma^2}{n\varepsilon^2} \to 0$ $(n \to \infty)$)

3.2.1 Csebisev-egyenlőtlenség

EX véges.
Ekkor $P(|X - EX| \geq \lambda) \leq \frac{D^2X}{\lambda^2}$
Megjegyzés: Bizonyítás Markov-egyenlőtlenséggel.

3.2.2 Markov-egyenlőtlenség

$X \geq 0, c > 0$.
Ekkor $P(X \geq c) \leq \frac{EX}{c}$

3.3 Konvergenciafajták

$\xi_n \to \xi$, vagyis ξ konvergens.

- *sztochasztikusan*: ha $\forall \varepsilon > 0$-ra $P(|\xi_n - \xi| \geq \varepsilon) \to 0$ $(n \to \infty)$.
- *1 valószínűséggel (majdnem mindenütt)*: ha $P(\omega : \xi_n(\omega) \to \xi(\omega)) = 1$.
- *L^p-ben*: ha $E(|\xi_n - \xi|^p) \to 0$ $(n \to \infty)$ $(p > 0$ rögzített).
- *eloszlásban*: ha $F_{\xi_n}(x) \to F_\xi(x)$ $(n \to \infty)$ az utóbbi minden folytonossági pontjában.

Kapcsolataik: 1 valószínűségű és L^p-beli a legerősebb, ezekből következik a sztochasztikus, ebből pedig az eloszlásbeli.

4 Centrális határeloszlás tétel

$X_1, X_2, ...$ függetlenek, azonos eloszlásúak, $EX_1 = m < \infty$, $D^2X_1 = \sigma^2 < \infty$.
Ekkor $\frac{X_1+...+X_n - nm}{\sqrt{n}\sigma} \to N(0,1)$ $(n \to \infty)$ eloszlásban, azaz $P(\frac{X_1+...+X_n-nm}{\sqrt{n}\sigma} < x) \to \Phi(x)$ $(n \to \infty)$.

5 Statisztikai mező

$(\Omega, \mathcal{A}, \mathcal{P})$ hármas, ha $\mathcal{P} = \{P_\vartheta\}_{\vartheta \in \Theta}$ és $(\Omega, \mathcal{A}, P_\vartheta)$ Kolmogorov-féle valószínűségi mező $\forall \vartheta \in \Theta$-ra.

5.1 Fogalmak

- *Minta*: $\underline{\xi} = (\xi_1, ..., \xi_n) : \Omega \to \Xi \in \mathbb{R}^n$. ($\xi_i$ valószínűségi változó)
- *Mintatér*: Ξ, minta lehetséges értékeinek halmaza, gyakran $\mathbb{R}^n, \mathbb{Z}^n$.
- *Minta [realizációja]*: $\underline{x} = (x_1, ..., x_n)$, konkrét megfigyelés.
- *Statisztika*: $T : \Xi \to \mathbb{R}^k$.
- *Statisztika alaptétele*: (Glivenko–Cantelli-tétel) $\xi_1, \xi_2, ...$ független, azonos eloszlású F eloszlásfüggvénnyel. Ekkor az F_n tapasztalati eloszlásfüggvényre teljesül, hogy $\sup_{-\infty<x<\infty} |F_n(x) - F(x)| \to 0$ $(n \to \infty)$ 1 valószínűséggel.

6 Statisztikai becslések

$(\Omega, \mathcal{A}, \mathcal{P})$ statisztikai mező, $\vartheta \in \Theta$, $P_\vartheta(\xi_1 < x_1, ..., \xi_n < x_n) = F_\vartheta(\underline{x})$
$T(\underline{\xi})$ a ϑ *becslése*, ha $T : \mathbb{R}^n \to \Theta$.
$T(\underline{\xi})$ a $h(\vartheta)$ *becslése*, ha $T : \mathbb{R}^n \to h(\Theta)$.
Torzítatlanság: $T(\underline{\xi})$ torzítatlan becslése $h(\vartheta)$-nak, ha $E_\vartheta T(\underline{\xi}) = h(\vartheta)$ $\forall\vartheta \in \Theta$.
Aszimptotikusan torzítatlan: $T(\underline{\xi})$ aszimptotikusan torzítatlan a $h(\vartheta)$-ra, ha $E_\vartheta T(\underline{\xi}) \to h(\vartheta)$ $(n \to \infty)$ $\forall\vartheta \in \Theta$.
A T_1 torzítatlan becslés *hatásosabb* T_2 torzítatlan becslésnél, ha $D^2_\vartheta T_1 \leq D^2_\vartheta T_2$ $\forall\vartheta \in \Theta$.
Hatásos, ha minden más torzítatlan becslésnél hatásosabb. Ha van hatásos becslés, akkor az egyértelmű.

- *Maximum-likelihood becslés*: Likelihood függvény: $L(\vartheta, \underline{x}) = \begin{cases} P_\vartheta(\underline{\xi} = \underline{x}) & diszkr. \\ f_{\vartheta,\underline{\xi}}(\underline{x}) & absz.folyt. \end{cases}$
  Független esetben: $L(\vartheta, \underline{x}) = \begin{cases} \prod_{i=1}^n P_\vartheta(\xi_i = x_i) & diszkr. \\ \prod_{i=1}^n f_{\vartheta,\xi_i}(x_i) & absz.folyt. \end{cases}$
  $\hat{\vartheta}$ a ϑ ismeretlen paraméter maximum-likelihood becslése, ha $L(\hat{\vartheta}, \underline{\xi}) = \max_{\vartheta\in\Theta} L(\vartheta, \underline{\xi})$.
- *Momentum-módszer becslés*: $\vartheta = (\vartheta_1, ..., \vartheta_k)$, $\xi_1, ..., \xi_n$, l. momentum: $M_l(\underline{\vartheta}) = E_{\underline{\vartheta}}\xi_i^l$, tapasztalati l. momentum: $\hat{M}_l = \frac{\sum_{i=1}^n \xi_i^l}{n}$.
  $\underline{\hat{\vartheta}}$ a $\underline{\vartheta}$ momentum módszer szerinti becslése, ha megoldása az $M_l(\underline{\vartheta}) = \hat{M}_l$, $l = 1..k$ egyenletrendszernek.

7 Hipotézisvizsgálat

Felteszünk egy hipotézist, és vizsgáljuk, hogy igaz-e. Elfogadjuk vagy elutasítjuk. Lehet paraméteres vagy nem paraméteres, vizsgálhatjuk várható értékek, szórások egyezőségét, értékét, teljes eseményrendszerek függetlenségét. Illeszkedésvizsgálattal megállapíthatjuk, hogy a valószínűségi változók adott eloszlásfüggvényűek-e, homogenitásvizsgálattal pedig azt, hogy ugyanolyan eloszlású-e két minta.
H_0: nullhipotézis, $\vartheta \in \Theta_0$; H_1: ellenhipotézis, $\vartheta \in \Theta_1$; $\Theta = \Theta_0 \bigcup \Theta_1$.
Egy- és kétoldali vizsgálat: Kétoldali ellenhipotézisnél a nem egyezőséget tesszük fel, egyoldalinál valamilyen relációt. Kétoldalinál a próba értékének abszolút értékét vizsgáljuk, hogy az elfogadási tartományon belül van-e, ekkor például az u-próbánál az adott hibaszázalékot meg kell felezni a számításhoz, hiszen a Φ függvény szimmetrikus az y tengelyre.

- *Statisztikai próba*: $\Xi = \Xi_e \bigcup \Xi_k$ (diszjunkt halmazok) elfogadási és kritikus tartomány. Ez a felbontás a statisztikai próba. Ha a megfigyelés eleme a kritikus tartománynak, akkor elutasítjuk a nullhipotézist, ha nem eleme, akkor elfogadjuk. $T(\underline{x}) = \begin{cases} 1 & x \in \Xi_k \\ 0 & otherwise \end{cases}$
- *Elsőfajú hiba*: H_0 igaz, de elutasítjuk. Valószínűsége: $P_\vartheta(\underline{\xi} \in \Xi_k)$, $\vartheta \in \Theta_0$.
- *Másodfajú hiba*: H_0 hamis, de elfogadjuk. Valószínűsége: $P_\vartheta(\underline{\xi} \notin \Xi_k)$, $\vartheta \in \Theta_1$.
  Az a cél, hogy ezek a hibák minél kisebbek legyenek. Egymás kárára javítható a két valószínűség, ha a megfigyelések száma rögzített.
- *Próba terjedelme*: α a próba terjedelme, ha $P_\vartheta(\underline{\xi} \in \Xi_k) \leq \alpha$, $\vartheta \in \Theta_0$.
  α a próba pontos terjedelme, ha $\sup_{\vartheta\in\Theta_0} P_\vartheta(\underline{\xi} \in \Xi_k) = \alpha$.

8 Klasszikus statisztikai próbák

- *u-próba*: Feltételezzük, hogy a minta normális eloszlású ($\xi_i \sim N(m, \sigma^2)$), $i = 1..n$, és hogy a szórás ismert.
  - *Egymintás*: A nullhipotézis az, hogy a várható érték megegyezik-e egy konkrét értékkel (m_0), másképpen fogalmazva azt vizsgáljuk, hogy a mintabeli átlag nem tér-e el szignifikánsan m_0-tól. Tehát $H_0 : m = m_0$, és kétoldali esetben $H_1 : m \neq m_0$, egyoldaliban pedig például $H_1 : m \geq m_0$ vagy $H_1 : m < m_0$.
    Az u-próba értéke: $u = \sqrt{n}\frac{\bar{\xi}-m_0}{\sigma}$. Ha igaz a nullhipotézis, akkor ez közel standard normális eloszlású.
    ε hibavalószínűséggel vizsgáljuk a hipotézist, ehhez szükségünk van a $\Phi(u_{1-\varepsilon}) = 1-\varepsilon$ értékre.
    Kétoldali esetben H_0-t elutasítjuk, ha $|u| > u_{1-\frac{\varepsilon}{2}}$, és elfogadjuk, ha $|u| \leq u_{1-\frac{\varepsilon}{2}}$.
    Egyoldali esetben $u > u_{1-\varepsilon}$ (jobb) és $u < u_{1-\varepsilon}$ (bal) esetét vizsgáljuk, ezen esetekben utasítjuk el H_0-t.
  - *Kétmintás*: Itt a feltételek a következők: $\xi_i \sim N(m_1, \sigma_1^2)$, $i = 1..n$ és $\eta_j \sim N(m_2, \sigma_2^2)$, $j = 1..m$. A szórások szintén ismertek. $H_0 : m_1 = m_2$, és $u = \frac{\bar{\xi}-\bar{\eta}}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}}$. $H_1 : m_1 > m_2$, ez a felső (jobb?) oldali, $H_1 : m_1 < m_2$ pedig az alsó (bal?) ellenhipotézis.
- *t-próba*: Ennél a próbánál nem ismert a szórás, viszont ugyanúgy normális eloszlást feltételezünk, mint az u-próbánál. $\xi_i \sim N(m, \sigma^2)$, $i = 1..n$.
  - *Egymintás*: $H_0 : m = m_0$. Ellenhipotézis az u-próbához hasonlóan. $t = \sqrt{n}\frac{\bar{\xi}-m_0}{\sqrt{\sigma_*^2}}$, ahol σ_*^2 a korrigált tapasztalati szórásnégyzet, amit a mintából számíthatunk ki. (Megjegyzés: n helyett $n-1$-gyel osztunk a képletben.) Ez az érték t-eloszlású H_0 esetén, ami $n-1$ szabadságfokú. Más néven szokás ezt a próbát Student-próbának is nevezni.
  - *Kétmintás*: $\xi_i \sim N(m_1, \sigma_1^2)$, $i = 1..n$ és $\eta_j \sim N(m_2, \sigma_2^2)$, $j = 1..m$. Ez esetben sem ismert a szórás, viszont feltételezzük, hogy a két minta szórása megegyezik. Ekkor $t_{n+m-2} = \sqrt{\frac{nm(n+m-2)}{n+m}} \frac{\bar{\xi}-\bar{\eta}}{\sqrt{\sum(\xi_i-\bar{\xi})^2+\sum(\eta_j-\bar{\eta})^2}}$. $n+m-2$ a próba szabadságfoka.
- *f-próba*: Két minta esetén használható. Ez a próba szórások egyezőségének vizsgálatára alkalmas, tehát itt $H_0 : \sigma_1 = \sigma_2$. Ha a két minta szórásnégyzete megegyezik, akkor a hányadosuk 1-hez tart. $f_{n-1,m-1} = max(\frac{\sigma_1^2}{\sigma_2^2}, \frac{\sigma_2^2}{\sigma_1^2})$. A két szabadsági fok közül az első az f számlálójához tartozó minta elemszáma -1, a második a nevezőjéhez.
- *Welch-próba*: Más néven d-próba. Hasonló, mint a kétmintás t-próba, de itt a szórások egyezőségét nem kell feltenni. Szabadsági foka bonyolult képlettel számítható.
- *szekvenciális próbák*: $V_n = \frac{\prod f_1(x_i)}{\prod f_0(x_i)} = \frac{L_1(\underline{x})}{L_0(\underline{x})}$. f_0 a nullhipotézis szerinti sűrűségfüggvény, f_1 az ellenhipotézis szerinti. Adott egy A és egy B érték, $A < B$. Ha $V_n \geq B$, akkor elutasítjuk H_0-t, ha $V_n \leq A$, akkor elfogadjuk, és ha $A < V_n < B$, akkor új mintaelemet veszünk.
  Stein tétele szerint N 1 valószínűséggel véges. $N = min\{n : V_n \leq A \vee V_n \geq B\}$.
- *Minőség-ellenőrzés*: n_1 elemet nézünk, $c_1 < c_2$ és c_3 határértékek. Ha $X_1 \leq c_1$, akkor elfogadjuk H_0-t, ha $X_1 \geq c_2$, akkor elutasítjuk. Ha $c_1 < X_1 < c_2$, akkor megnézünk n_2 elemet, és ha $X_1 + X_2 \leq c_3$, akkor szintén elfogadjuk H_0-t. A várható mintaelemszám méri a hatékonyságát.
- χ^2-*próba*: $H_0 : A_1, ..., A_n$ teljes eseményrendszer. $P(A_i) = p_i, i = 1..n$, ν_i a gyakoriság. Ha teljesül a nullhipotézis, akkor $\frac{\nu_i}{n} \sim p_i$.
  $\chi^2 = \sum \frac{(\nu_i - np_i)^2}{np_i}$. Ez χ^2 eloszlású, aminek $r-1$ szabadságfoka van. r az összeadott csoportok száma. (Megjegyzés: ha túl kicsi lenne 1-1 csoportban a gyakoriság, akkor azokat összevonjuk.)
  χ^2-próbát használhatunk illeszkedés-, homogenitás- és függetlenségvizsgálatra is. (Megjegyzés: más képlet van mindhez.)
- *Egyéb próbák*:
  - *Kolmogorov–Szmirnov-próba*: 2 tapasztalati eloszlásfüggvény megegyezik-e (homogenitásvizsgálat), vagy 1 minta esetén megegyezik-e valamilyen eloszlásfüggvénnyel. $D_{m,n} = \max_x |F_n(x) - G_m(x)|$. X_i F eloszlásfüggvénnyel, Y_j G-vel. $H_0 : F \equiv G$.

- *Előjel-próba*: Hányszor teljesül, hogy valami pozitív.
- *Wilcoxon-próba*: (rangstatisztika), $P(X > Y) = \frac{1}{2}$ tesztelésére összeszámoljuk, hogy hány párra teljesül, hogy $X_i > Y_j$.