Segédanyag a Statisztika (alk.mat.) tantárgyhoz

2015. május 5.

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

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

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

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

- $E(X+Y) = EX + EY$;
- $E(aX) = aEX$;
- $E\sum_{i=1}^{n} c_iX_i = \sum_{i=1}^{n} c_iEX_i$;
- $D^2(aX+b) = a^2D^2X$.

Állítás. Tetszőleges X val. változó esetén

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

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

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	Exp(λ)	$\begin{cases}1-e^{-\lambda x} & \text{ha } x \ge 0\\ 0 & \text{különben}\end{cases}$	$\begin{cases}\lambda e^{-\lambda x} & \text{ha } x \ge 0\\ 0 & \text{különben}\end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Standard normális	N($0,1^2$)	$\Phi(x) = ...$	$\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \quad x \in \mathbb{R}$	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

További nevezetes abszolút folytonos eloszlások:

Eloszlás neve	Jelölése	Eloszlásfüggvény	Sűrűségfüggvény	EX	D^2X
Cauchy	$Cauchy(a,b)$ $a \in \mathbb{R}, b > 0$	$\frac{1}{\pi}\text{arctg}\left(\frac{x-a}{b}\right) + \frac{1}{2}$	$\frac{1}{\pi b\left[1+\left(\frac{x-a}{b}\right)^2\right]} \quad x \in \mathbb{R}$	∄	∄
Pareto*	$Pareto(\alpha,\beta)$ $\alpha, \beta > 0$	$\begin{cases}1-\left(\frac{\beta}{x}\right)^{\alpha} & \text{ha } x \ge \beta\\ 0 & \text{ha } x < \beta\end{cases}$	$\begin{cases}\frac{\alpha}{\beta}\left(\frac{\beta}{x}\right)^{\alpha+1} & \text{ha } x \ge \beta\\ 0 & \text{ha } x < \beta\end{cases}$	$\frac{\alpha\beta}{\alpha-1}$	$\frac{\beta^2\alpha}{(\alpha-1)^2(\alpha-2)}$

Eloszlás neve	Jelölése	Eloszlás-függvény	Sűrűségfüggvény	EX	D^2X
Khí-négyzet	$\chi^2_k \quad k \in \mathbb{N}$	...	$\frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}e^{-x/2} \quad x \in \mathbb{R}$	k	$2k$
Gamma	$\Gamma(\alpha,\lambda)$ $\alpha, \lambda > 0$	...	$\begin{cases}\frac{1}{\Gamma(\alpha)}\lambda^{\alpha}e^{-\lambda x}x^{\alpha-1} & \text{ha } x \ge 0\\ 0 & \text{ha } x < 0\end{cases}$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
Béta	$Beta(\alpha,\beta)$ $\alpha, \beta > 0$	...	$\begin{cases}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} & x \in [0,1]\\ 0 & \text{különben}\end{cases}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Lognormális	LN(m,σ^2) $m \in \mathbb{R}, \sigma > 0$	...	$\begin{cases}\frac{1}{x\sqrt{2\pi}\sigma}e^{-\frac{(\log x-m)^2}{2\sigma^2}} & \text{ha } x \le 0\\ 0 & \text{ha} x < 0\end{cases}$	$e^{m+\sigma^2/2}$	$(e^{\sigma^2-1})e^{2m+\sigma^2}$

* A Pareto-eloszlásnak akkor van véges várható értéke a képletnek megfelelően, ha $\alpha > 1$, szórásnégyzete pedig akkor, ha $\alpha > 2$.

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

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

Állítás. Legyenek $X_1, ..., X_n$, X és Y független val. változók

- $X_1 \sim \text{Ind}(p), ..., X_n \sim \text{Ind}(p) \quad \Rightarrow \quad X_1 + ... + X_n \sim \text{Bin}(n,p)$
- $X \sim \text{Bin}(n,p)$, $Y \sim \text{Bin}(m,p) \quad \Rightarrow \quad X + Y \sim \text{Bin}(n+m,p)$

- $X_1 \sim \text{Geo}(p), ..., X_n \sim \text{Geo}(p) \quad \Rightarrow \quad X_1 + ... + X_n \sim \text{NegBin}(n, p)$
- $X \sim \text{Poi}(\lambda_1), Y \sim \text{Poi}(\lambda_2) \quad \Rightarrow \quad X + Y \sim \text{Poi}(\lambda_1 + \lambda_2)$
- $X \sim N(m_1, \sigma_1^2), Y \sim N(m_2, \sigma_2^2) \quad \Rightarrow \quad X+Y \sim N(m_1+m_2, \sigma_1^2+\sigma_2^2)$
- $X_1 \sim \text{Exp}(\lambda), ..., X_n \sim \text{Exp}(\lambda) \quad \Rightarrow \quad X_1 + ... + X_n \sim \Gamma(n, \lambda)$
- $X \sim \Gamma(\alpha, \lambda), Y \sim \Gamma(\beta, \lambda) \quad \Rightarrow \quad X + Y \sim \Gamma(\alpha + \beta, \lambda)$

Tétel. **Valószínűségi vektorváltozó transzformáltjának sűrűségfüggvénye.** Legyen $\underline{X} = (X_1, \dots, X_n)$ abszolút folytonos valószínűségi vektorváltozó $f_{\underline{X}}$ sűrűségfüggvénnyel, $A \subseteq \mathbb{R}^n$ összefüggő és nyílt halmaz. Legyen $\underline{g} : A \to D \subset \mathbb{R}^n$ függvény, amely invertálható és inverze folytonosan differenciálható. Legyen $\underline{Y} = \underline{g}(\underline{X}), \quad J = \partial_{\underline{y}} \underline{g}^{-1}(\underline{y})$ a Jacobi-mátrix. Ekkor $f_{\underline{g}(\underline{X})}(\underline{y}) = |\det(J)| \cdot f_{\underline{X}}(\underline{g}^{-1}(\underline{y}))$

Definíció. **Kovarianciamátrix.** Legyen $\underline{X}$ valószínűségi vektorváltozó. Ekkor $\Sigma := E(\underline{X} \cdot \underline{X}^T) - E(\underline{X})E(\underline{X})^T$

Legyen $X_1, X_2, \dots$ valószínűségi változók sorozata.

Definíció. **1 valószínűségű konvergencia:** (jel.: $X_n \xrightarrow[n\to\infty]{\text{1 vsz.}} X$)

$X_n \overset{n\to\infty}{\longrightarrow} X$ *1 valószínűséggel*, ha $P(\{\omega : X_n(\omega) \overset{n\to\infty}{\to} X(\omega)\}) = 1$.

Definíció. **sztochasztikus konvergencia:** (jel.: $X_n \xrightarrow[n\to\infty]{p} X$)

$X_n \overset{n\to\infty}{\longrightarrow} X$ *sztochasztikusan*, ha $\forall \varepsilon > 0$-ra $\lim\limits_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$.

Definíció. **eloszlásbeli konvergencia:** (jel.: $X_n \xrightarrow[n\to\infty]{d} X$)

$X_n \overset{n\to\infty}{\longrightarrow} X$ *eloszlásban*, ha $F_{X_n}(x) \overset{n\to\infty}{\longrightarrow} F_X(x)$ minden x folytonossági pontban.

Definíció. **L^p-beli konvergencia:** (jel.: $X_n \xrightarrow[n\to\infty]{L^p} X$)

$X_n \overset{n\to\infty}{\longrightarrow} X$ *L^p-ben*, ha $E\left(|X_n - X|^p\right) \overset{n\to\infty}{\longrightarrow} 0$.

Minta: $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$ valószínűségi változó sorozat. A továbbiakban feltesszük, hogy függetlenek és azonos eloszlásúak – *i.i.d. minta*.
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.

Definíció. **kvantilisfüggvény:** $F^{-1}(y) = q(y) = q_y = \inf\{x : F(x) \geq y\}$, ahol $0 < y < 1$
Fontos speciális kvantilisek: kvartilisek:

- $Q_1 := q_{\frac{1}{4}} \rightsquigarrow$ alsó kvartilis
- $Q_2 = Me := q_{\frac{1}{2}} \rightsquigarrow$ **medián**
- $Q_3 := q_{\frac{3}{4}} \rightsquigarrow$ felső kvartilis

Néhány lényeges statisztika:

- *Rendezett minta*: $X_1^* \leq \dots \leq X_n^*$ nem csökkenő sorrendbe tesszük a mintaelemeket
- *Mintaátlag*: $\overline{X} = \frac{\sum\limits_{i=1}^{n} X_i}{n}$
- *Tapasztalati szórásnégyzet*: $S_n^2 = \frac{\sum\limits_{i=1}^{n} (X_i - \overline{X})^2}{n}$
- *Korrigált tapasztalati szórásnégyzet*: $(S_n^*)^2 = \frac{\sum\limits_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$
- *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
- *Módusz:* abszolút folytonos eloszlás esetén a sűrűségfüggvény maximumhelye(i), diszkrét eloszlás esetén pedig az eloszlás maximumhelye(i)

Tétel. **(Glivenko-Cantelli)** A tapasztalati eloszlásfüggvény 1 valószínűséggel egyenletesen tart a valódi eloszlásfüggvényhez, formálisan $P\left(\lim\limits_{n\to\infty} \sup\limits_{x\in\mathbb{R}} |F_n(x) - F(x)| = 0\right) = 1$.

Állítás. Legyen $(X_1, \dots, X_n)$ i.i.d. minta egy abszolút folytonos eloszlásból. Jelölje $F(x)$ a közös eloszlásfüggvényt, $f(x)$ pedig a közös sűrűségfüggvényt. Ekkor a k-adik rendezett mintaelem sűrűségfüggvénye $f_{X_k^*} = \frac{n!}{(k-1)!(n-k)!} f(x)[F(x)]^{k-1}[1 - F(x)]^{n-k}$.
Speciálisan,

$k=1$ esetén $\quad F_{X_1^*}=[1-F(x)]^n$ és $\quad f_{X_1^*}=nf(x)[1-F(x)]^{n-1}$
$k=n$ esetén $\quad F_{X_n^*}=[F(x)]^n$ és $\quad f_{X_n^*}=nf(x)[F(x)]^{n-1}$

Definíció. **Statisztikai mező.** $(\Omega,\mathcal{A},\mathcal{P})$ hármas, ahol $(\Omega,\mathcal{A})$ mérhető tér, $\mathcal{P}$ pedig eloszlások egy családja.
$\mathcal{P}$-t gyakran paraméresen adjuk meg: $\mathcal{P}=\{P_\vartheta:\ \vartheta\in\Theta\}$, ahol $\Theta\subseteq\mathbb{R}^p$ összefüggő és nyílt halmaz, amit **paramétertér**nek hívunk.

Definíció. **Minta.** $\mathbf{X}:(\Omega,\mathcal{A})\longrightarrow(\mathcal{X},\mathcal{B})$ mérhető leképezés, ahol $(\mathcal{X},\mathcal{B})$ neve: **mintatér** (néha csak $\mathcal{X}$-et hívják mintatérnek).
A minta koordinátái jellemzően függetlenek és azonos eloszlásúak, azaz $\mathbf{X}=(X_1,X_2,\dots,X_n)^T$.

Feladat: annak a meghatározása, hogy a $\mathcal{P}$ eloszláscsalád melyik tagja írja le legjobban a valóságot, a vizsgált jelenséget. Ennek érdekében veszünk mintát.

Definíció. **Likelihood függvény:** Legyen $\mathbf{X}=(X_1,...,X_n)$ i.i.d. minta

- $L(\vartheta;\mathbf{x})=f_\vartheta(\mathbf{x})=\prod_{i=1}^n f_\vartheta(x_i)$, ha az eloszlás folytonos
- $L(\vartheta;\mathbf{x})=P_\vartheta(\mathbf{X}=\mathbf{x})=\prod_{i=1}^n P_\vartheta(X_i=x_i)$, ha az eloszlás diszkrét.

Definíció. **Elégséges statisztika.**
Legyen $(\Omega,\mathcal{A},\mathcal{P})$ statisztikai mező, $\mathbf{X}$ minta, $B\in\mathcal{A}$.
A T statisztikát elégséges statisztikának nevezzük, ha a $P_\vartheta(\mathbf{X}\in B|T(\mathbf{X}))$ feltételes eloszlásnak létezik ϑ-tól nem függő változata.
Megj.: az elégséges statisztika minden lényeges információt tartalmaz az ismeretlen ϑ paraméterre vonatkozóan. Szeretnénk, ha ez minél "gazdaságosabb", tömörebb lenne $\longrightarrow$ ennek a fogalmát ragadja meg a minimális elégséges statisztika.

Definíció. **Minimális elégséges statisztika.**
A T elégséges statisztika minimális elégséges, ha minden S elégséges statisztikához létezik olyan φ mérhető függvény, hogy $T=\varphi(S)$.

Definíció. **Dominált statisztikai mező.**
Az $(\Omega,\mathcal{A},\mathcal{P})$ statisztikai mezőt dominánltnak nevezzük, ha létezik olyan $\lambda:\mathcal{A}\to\overline{\mathbb{R}}$ σ-véges mérték, hogy minden $P\in\mathcal{P}$ esetén $P<<\lambda$.

Tétel. **Neyman-féle faktorizációs tétel.**
Dominált statisztikai mezőn a T statisztika akkor és csak akkor elégséges, ha léteznek olyan g_ϑ nemnegatív és h mérhető függvények, hogy $L(\vartheta;\mathbf{x})=g_\vartheta(T(\mathbf{x}))\cdot h(\mathbf{x})\quad\forall\vartheta\in\Theta$ és λ-m.m. $\mathbf{x}\in\mathcal{X}$ esetén.

Tétel.
a.) T elégséges $\Longleftrightarrow$
$\Longleftrightarrow$ (λ-m.m. $\mathbf{x},\mathbf{y}\in\mathcal{X}$ esetén $T(\mathbf{x})=T(\mathbf{y})\ \Rightarrow\ \frac{L(\vartheta;\mathbf{x})}{L(\vartheta;\mathbf{y})}$ nem függ ϑ-tól)
b.) T minimális elégséges $\Longleftarrow$
$\Longleftarrow$ (λ-m.m. $\mathbf{x},\mathbf{y}\in\mathcal{X}$ esetén $T(\mathbf{x})=T(\mathbf{y})\ \Leftrightarrow\ \frac{L(\vartheta;\mathbf{x})}{L(\vartheta;\mathbf{y})}$ nem függ ϑ-tól)

Állítás. A $T(\mathbf{X})=\mathbf{X}^*$ rendezett minta elégséges statisztika.

Legyen $g:\Theta\to\mathbb{R}^k$ függvény. Célunk az $\mathbf{X}$ minta alapján $g(\vartheta)$ becslése.

Definíció. **Torzítatlan becslés.**
T($\mathbf{X}$) stat. torzítatlan becslése $g(\vartheta)$-nak, ha $E_\vartheta T(\mathbf{X})=g(\vartheta)\quad\forall\vartheta\in\Theta$-ra.

Definíció. **Torzítás (bias).** $b_T(\vartheta)=E_\vartheta T(\mathbf{X})-g(\vartheta)$

Definíció. Legyenek $T_1(\mathbf{X})$ és $T_2(\mathbf{X})$ torzítatlan becslései $g(\vartheta)$-nak. Ekkor azt mondjuk, hogy $T_1(\mathbf{X})$ **hatásosabb** $T_2(\mathbf{X})$-nél, ha $D^2_\vartheta(T_1(\mathbf{X}))\le D^2_\vartheta(T_2(\mathbf{X}))$ minden $\vartheta\in\Theta$ esetén.

Definíció. **Hatásos becslés.**
A $T(\mathbf{X})$ tozítatlan becslést hatásosnak nevezzük, ha minden torzítatlan becslésnél hatásosabb.

Tétel. **A hatásos becslés egyértelműsége.**
Ha $T_1(\mathbf{X})$ és $T_2(\mathbf{X})$ hatásos becslései $g(\vartheta)$-nak, akkor minden paraméterértékre 1 valószínűséggel megegyeznek, azaz $P_\vartheta(T_1(\mathbf{X})=T_2(\mathbf{X}))=1\ \ \forall\vartheta\in\Theta$ esetén.

Definíció. **Aszimptotikus torzítatlanság.**
A $T_n(\mathbf{X})$ becléssorozat $(n=1,2,...)$ aszimptotikusan torzítatlan becslése a $g(\vartheta)$-nak, ha $E_\vartheta T_n(\mathbf{X})\xrightarrow{n\to\infty}g(\vartheta)\ \ \forall\vartheta\in\Theta$ esetén.

Definíció. **Gyenge konzisztencia.**
A $T_n(\mathbf{X})$ becléssorozat $(n=1,2,...)$ gyengén konzisztens becslése a $g(\vartheta)$-nak , ha $T_n(\mathbf{X})\xrightarrow[n\to\infty]{p}g(\vartheta)\ \ \forall\vartheta\in\Theta$ esetén.

Tétel. **Elégséges feltétel gyenge konzisztenciára.**

Ha $E_\vartheta T_n(\mathbf{X}) \xrightarrow{n\to\infty} g(\vartheta)$ és $D^2_\vartheta T_n(\mathbf{X}) \xrightarrow{n\to\infty} 0$, akkor T_n becsléssorozat gyengén konzisztens becslése $g(\vartheta)$-nak.

Definíció. **Erős konzisztencia.**
A $T_n(\mathbf{X})$ becsléssorozat ($n = 1, 2, ...$) erősen konzisztens becslése a $g(\vartheta)$-nak, ha $T_n(\mathbf{X}) \xrightarrow[n\to\infty]{1 \text{ vsz.}} g(\vartheta)$ $\forall\vartheta \in \Theta$ esetén.

Legyen a mintatér p dimenziós, azaz $\Theta \subset \mathbb{R}^p$. A likelihood függvény $\forall\vartheta$-ra 1 valószínűséggel pozitív, így lehet venni a logaritmusát.

Definíció. **Log-likelihood függvény.** $l(\vartheta; \mathbf{x}) = \log L(\vartheta; \mathbf{x})$
A továbbiakban a gradienset mindig sorvektornak tekintjük:
$\partial_\vartheta l(\vartheta; \mathbf{x}) = \big(\partial_{\vartheta_1} l(\vartheta; \mathbf{x}), \ldots, \partial_{\vartheta_p} l(\vartheta; \mathbf{x})\big)$

Definíció. **Fisher-információ.**
Tegyük fel, hogy m.m. $\mathbf{x} \in \mathcal{X}$-re a log-likelihood függvény ϑ szerint deriválható. Ekkor az $\mathbf{X}$ n elemű mintában lévő Fisher-információ:
$I_\mathbf{X}(\vartheta) \equiv I_n(\vartheta) = E_\vartheta\left([\partial_\vartheta l(\vartheta; \mathbf{X}]^T [\partial_\vartheta l(\vartheta; \mathbf{X}]\right)$.

Megj.: $I_\mathbf{X}(\vartheta)$ azt az (absztrakt) információmennyiséget méri, amelyet az $\mathbf{X}$ minta a paraméterre vonatkozóan magában hordoz.

Definíció. **Statisztika Fisher-információja.**
$I_T(\vartheta) \equiv I_{T(\mathbf{X})}(\vartheta) = E_\vartheta\left([\partial_\vartheta l(\vartheta; T(\mathbf{X})]^T [\partial_\vartheta l(\vartheta; T(\mathbf{X})]\right)$

Állítás. A Fisher-információ

- $p \times p$-es mátrix;
- szimmetrikus;
- pozitív szemidefinit.

A Fisher-információ kiszámítása bizonyos, úgynevezett regularitási feltételek esetén egyszerűbbé válik.

Definíció. **(R): gyenge regularitási feltételek.**

- $\sqrt{L(\vartheta; \mathbf{x})}$ folytonosan deriválható ϑ szerint m.m. $\mathbf{x} \in \mathcal{X}$-re
- $I_\mathbf{X}(\vartheta)$ létezik, véges, pozitív definit és folytonos

Definíció. **1. regularitási feltétel.**

- $E_\vartheta(\partial_\vartheta l(\vartheta, \mathbf{X})) = \mathbf{0}$

Állítás. $E_\vartheta(\partial_\vartheta l(\vartheta, \mathbf{X})) = \mathbf{0} \iff \partial_\vartheta \int\limits_{\mathbf{x}\in\mathcal{X}} f_\vartheta(\mathbf{x})\,\lambda(d\mathbf{x}) = \int\limits_{\mathbf{x}\in\mathcal{X}} \partial_\vartheta f_\vartheta(\mathbf{x})\,\lambda(d\mathbf{x})$, azaz "be lehet deriválni" az integráljel mögé.

Tétel. Tegyük fel, hogy (R) teljesül. Legyen $T : \mathcal{X} \to \mathbb{R}^k$ olyan statisztika, amelyre $E_\vartheta\left(\| T(\mathbf{X}) \|^2\right)$ a paramétertér minden pontjának egy környezetében korlátos (lokálisan korlátos). Ekkor $E_\vartheta(T(\mathbf{X})) = \int\limits_\mathcal{X} [T(\mathbf{x}) f_\vartheta(\mathbf{x})]\,\lambda(d\mathbf{x})$ folytonosan deriválható ϑ szerint, és "be lehet deriválni" az integráljel mögé, azaz $\partial_\vartheta E_\vartheta(T(\mathbf{X})) = \int\limits_{\mathbf{x}\in\mathcal{X}} [T(\mathbf{x}) \cdot \partial_\vartheta f_\vartheta(\mathbf{x})]\,\lambda(d\mathbf{x})$.

Következmény. $\partial_\vartheta E_\vartheta(T(\mathbf{X})) = E_\vartheta[T(\mathbf{X})\partial_\vartheta l(\vartheta, \mathbf{X})]$

Következmény. Speciálisan, ha $T(\mathbf{X}) = 1$, akkor $0 = E_\vartheta[T(\mathbf{X})\partial_\vartheta l(\vartheta, \mathbf{X})]$, azaz teljesül az 1. regularitási feltétel.

Állítás. Ha teljesül az 1. regularitási feltétel, akkor $I_n(\vartheta) = \Sigma_\vartheta(\partial_\vartheta l(\vartheta, \mathbf{X})^T)$, ahol $\Sigma(\mathbf{Y})$ az $\mathbf{Y}$ valószínűségi vektorváltozó kovariancia mátrixát jelöli.

Tétel. Tegyük fel, hogy $\mathbf{X}$ és $\mathbf{Y}$ egymástól független i.i.d. minták és a T statisztika paraméteres eloszláscsaládjára teljesül (R), és az alap valószínűségi tér eloszláscsaládjának minden elemének ugyanaz a tartója. Ekkor

a.) $I_{(\mathbf{X},\mathbf{Y})}(\vartheta) = I_\mathbf{X}(\vartheta) + I_\mathbf{Y}(\vartheta)$, azaz független minták infója összeadódik;
b.) $I_\mathbf{X}(\vartheta) \geq I_{T(\mathbf{X})}(\vartheta)$, azaz a statisztikában lévő infó nem lehet több, mint a mintában lévő infó;
c.) T elégséges $\iff I_\mathbf{X}(\vartheta) = I_{T(\mathbf{X})}(\vartheta)$, azaz az elégséges statisztika megőrzi az infót.

Következmény. Az előző tétel (a.) részéből következik, hogy amennyiben egy mintaelem eloszlására teljesül (R), akkor $I_n(\vartheta) = nI_1(\vartheta)$.

Tétel. **Fisher-információ átparaméterezés esetén.**
Legyen $\mathcal{T} \subset \mathbb{R}^q$ összefüggő és nyílt halmaz, $h : \mathcal{T} \to \Theta \subset \mathbb{R}^p$ folytonosan deriválható, injektív függvény. Tekintsük a $\mathcal{P}' = \{\mathcal{P}_{h(t)} : t \in \mathcal{T}\}$. Jelölje $I(t)$ a Fisher-információt a $\mathcal{P}'$ téren. Ekkor
$I(t) = h'(t)^T \cdot I(h(t)) \cdot h'(t)$, ahol $h'(t)$ $p \times q$-as mátrix.

Tétel. **Cramér-Rao egyenlőtlenség.**
Tegyük fel, hogy $T(\mathbf{X})$ statisztika torzítatlan becslése $g(\vartheta)$-nak és teljesül (R). Ekkor minden $\vartheta \in \Theta$-ra $\Sigma_\vartheta(T(\mathbf{X})) \geq \underbrace{g'(\vartheta) \cdot I_n(\vartheta)^{-1} \cdot g'(\vartheta)^T}_{\text{neve: információs határ}}$.
Ha minden ϑ-ra egyenlőség teljesül, akkor T hatásos becslés.

Definíció. **Teljes statisztika.** A T statisztika teljes, ha tetszőleges h

mérhető függvényére $\quad E_\vartheta(h(S)) = 0 \ \ \forall\vartheta\text{-ra} \iff h(S) = 0 \quad \mathcal{P}\text{-m.m.}$

Tétel. **Blackwell-Rao tétel.**
Tegyük fel, hogy T statisztika torzítatlan becslése $g(\vartheta)$-nak és S elégséges statisztika. Ekkor $E(T|S)$ feltételes várható érték torzítatlan becslése $g(\vartheta)$-nak és $\Sigma_\vartheta(E(T|S)) \leq \Sigma_\vartheta(T)$ minden ϑ-ra.
Ha S még teljes is, akkor $E(T|S)$ hatásos becslés.
Az eljárás neve: *blackwellizálás.*

Paraméterbecslési módszerek

- **Maximum likelihood módszer (ML-módszer):** Azt a paraméterértéket keressük, ahol a likelihood függvény a legnagyobb értéket veszi fel: $\max\limits_{\vartheta\in\Theta} L(\vartheta, \mathbf{x})$
  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: $\partial_\vartheta l(\vartheta, \mathbf{x}) = 0 \quad \rightsquigarrow \quad \hat{\vartheta}$
  - másodrendű feltétel: $\partial^2_\vartheta l(\hat{\vartheta}, \mathbf{x}) < 0$

  Ha ϑ p dimenziós, akkor $\vartheta = (\vartheta_1, ..., \vartheta_p)$, az
  - elsőrendű feltétel:
    $\partial_{\vartheta_i} l(\vartheta, \mathbf{x}) = 0 \quad \rightsquigarrow \quad \hat{\vartheta}_i \ \ (i = 1, ..., p) \quad \rightsquigarrow \quad \hat{\vartheta} = (\hat{\vartheta}_1, ..., \hat{\vartheta}_p)$
  - másodrendű feltétel: $H(\hat{\vartheta}_1, ..., \hat{\vartheta}_p) = \left(\partial_{\vartheta_i}\partial_{\vartheta_j} l(\hat{\vartheta}, \mathbf{x})\right)_{i,j=1,...,p}$
    Hesse-mátrix negatív definit

  Tétel: Ha ϑ ML-becslése $\hat{\vartheta}$, akkor tetszőleges g mérhető függvény esetén $g(\vartheta)$ ML-becslése $g(\hat{\vartheta})$.
- **Momentum-módszer** (klasszikus)**:** A mintából számítható tapasztalati momentumokat ($m_i := \frac{\sum_j x_j^i}{n}$) egyenlővé tesszük az elméleti momentumokkal ($M_i := E_\vartheta X^i$), az elsőtől kezdve, mégpedig annyit, amennyi paraméter van. Csak azokkal a momentumokkal érdemes foglalkozni, amelyeknél az elméleti momentum függ a paramétertől. Tehát p darab ismeretlen paraméter esetén a következő p ismeretlenes egyenletrendszert oldjuk meg:

  $$\begin{aligned} M_1 &= m_1 \\ &\vdots \\ M_p &= m_p \end{aligned}$$

  Megjegyzés: $m_1 = \overline{x}$
- **Bayes-becslés:** A becsülendő ϑ ismeretlen paraméterről előzetesen rendelkezünk bizonyos információkkal, amiket fel szeretnénk használni: feltesszük, hogy ϑ eloszlása valamilyen ismert eloszlásból származik.
  Jelölések:
  - ϑ val. változó által felvett konkrét értékek: t
  - ϑ eloszlása: $Q \rightsquigarrow$ *apriori eloszlás*
  - ϑ sűrűségfüggvénye: $q(t)$
  - $\mathbf{X}$ feltételes eloszlása $\vartheta = t$ mellett: P_t
  - $f_t = \frac{dP_t}{d\lambda}$, ahol λ domináló mérték
  - $\mathbf{X}$ feltétel nélküli eloszlása: P_Q
  - ϑ feltételes eloszlása $\mathbf{X} = \mathbf{x}$ mellett: $Q^*(\cdot|\mathbf{x}) \rightsquigarrow$ *aposteriori eloszlás*
  - $\vartheta|\mathbf{X}$ feltételes sűrűségfüggvénye: $q^*(t|\mathbf{x})$

  Bayes-becslés: $\hat{\vartheta}(\mathbf{X}) = E(\vartheta|\mathbf{X})$ feltételes várható érték.
  Ennek kiszámítása abszolút folytonos esetben:

  $$E(\vartheta|\mathbf{x}) = \int\limits_{t\in\Theta} t f_{\vartheta|\mathbf{X}}(t|\mathbf{x})\, dt = \int\limits_{t\in\Theta} t \frac{f_{\vartheta,\mathbf{X}}(t,\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\, dt = \frac{\int\limits_{t\in\Theta} t f_{\vartheta,\mathbf{X}}(t,\mathbf{x})\, dt}{f_{\mathbf{X}}(\mathbf{x})} =$$

  $$= \frac{\int\limits_{t\in\Theta} t f_{\mathbf{X}|\vartheta}(\mathbf{x}|t) f_\vartheta(t)\, dt}{\int\limits_{t\in\Theta} f_{\mathbf{X}|\vartheta}(\mathbf{x}|t) f_\vartheta(t)\, dt}, \text{ amit átírva a fenti jelölésekkel:}$$

  $$E(\vartheta|\mathbf{x}) = \int\limits_{t\ \in\Theta} t q^*(t|\mathbf{x})\, dt = \frac{\int\limits_{t\in\Theta} t f_t(\mathbf{x}) q(t)\, dt}{\int\limits_{t\in\Theta} f_t(\mathbf{x}) q(t)\, dt} \text{ adódik.}$$

  Kiszámítása általános esetben:

  $$E(\vartheta|\mathbf{x}) = \int\limits_{t\in\Theta} t\, Q^*(dt|\mathbf{x}) = \int\limits_{t\in\Theta} \frac{t f_t(\mathbf{x})}{f_Q(\mathbf{x})} Q(dt) = \frac{\int\limits_{t\in\Theta} t f_t(\mathbf{x}) Q(dt)}{\int\limits_{t\in\Theta} f_t(\mathbf{x}) Q(dt)}$$

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:
$H_0:\ \theta \in \Theta_0$ $\rightsquigarrow$ nullhipotézis
$\underline{H_1:\ \theta \in \Theta_1}$ $\rightsquigarrow$ 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: $\mathcal{X} \to \mathbb{R}$, $\mathcal{X}_k = \{\underline{x} \in \mathcal{X} : \mathrm{T}(\underline{x}) > c\}$ c neve: kritikus érték
$\mathcal{X}_e = \{\underline{x} \in \mathcal{X} : \mathrm{T}(\underline{x}) \le c\}$

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: $\alpha = \sup\{\alpha(\theta)\colon\ \theta \in \Theta_0\}$

Azt mondjuk, hogy az 1-es próba *erősebb* a 2-es próbánál, ha $\alpha_1 = \alpha_2$ és $\psi_1(\theta) \ge \psi_2(\theta)\ \forall \theta \in \Theta_1$.
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 : $\mathrm{T}(\mathbf{x})= c_\alpha$.
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.

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.

$H_0:\ f = f_0$
$\underline{H_1:\ f = f_1}$
A valószínűség-hányados próba kritikus tartománya: $\mathcal{X}_k = \left\{\mathbf{x} : \frac{f_1(\underline{x})}{f_0(\underline{x})} > c_\alpha\right\}$
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.

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

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

1.) Egymintás próbák

a.) Egymintás u-próba

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

a.) $H_0 : m = m_0$, $\underline{H_1 : m \ne m_0}$ b.) $H_0 : m = m_0$, $\underline{H_1 : m > m_0}$ c.) $H_0 : m = m_0$, $\underline{H_1 : m < m_0}$

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

$X_1, \dots, X_n \sim \mathrm{N}(m, \sigma^2)$, 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

$X_1, \dots, X_n \sim \mathrm{N}(m_1, \sigma_1^2)$

$Y_1, \dots, Y_m \sim \mathrm{N}(m_2, \sigma_2^2)$

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

$m_1, m_2, \sigma_1, \sigma_2$ paraméterekek

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

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

A kritikus tartományt $s_1^* > s_2^*$ esetén általában a következő alakban állapítják meg: $\mathcal{X}_k = \{(\mathbf{X}, \mathbf{Y}) : F > F_{n-1,m-1,1-\alpha/2} \text{ vagy } F < F_{n-1,m-1,\alpha/2}\}$

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

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

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

$\frac{1}{f} = \frac{c^2}{n-1} + \frac{(1-c)^2}{m-1}$, ahol $c = \frac{\frac{(s_1^*)^2}{n}}{\frac{(s_1^*)^2}{n}+\frac{(s_2^*)^2}{m}}$, ha $s_1^* > s_2^*$

Definíció. **F-eloszlás:** Az X valószínűségi változó n és m szabadságfokú F-eloszlást követ (jel.: $X \sim F_{n,m}$), ha $X = \frac{\frac{Y_n}{n}}{\frac{Z_m}{m}}$, ahol $Y_n \sim \chi_n^2$ és $Z_n \sim \chi_m^2$ függetlenek egymástól.

Jel.: $F_{n,m;\alpha}$: n, m szabadságfokú F-eloszlás α-kvantilise

Állítás. $F_{n,m;\alpha} = \frac{1}{F_{m,n;1-\alpha}}$

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: $\mathrm{P}(X_1 = x_j) = p_j \quad j = 1, \dots, r$. Jelöljük N_j-vel a gyakoriságokat, azaz azt, hogy az n elemű mintában hány darab x_j szerepel. Formálisan $N_j = \sum_{i=1}^{n} I(X_i = x_j)$. Ennek tapasztalati verzióját n_j-vel jelöljük.

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: $\mathbf{p}{=}(p_1, \dots, p_r)$

H_1: nem ezek a valószínűségek

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

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

Becsléses illeszkedésvizsgálat: csak annyit "sejtünk", hogy a minta valami-

lyen eloszlású, viszont a paramétereiről nincs sejtésünk. Ilyenkor amennyiben ML-módszerrel becsüljük meg az s darab ismeretlen paramétert, akkor a próbastatisztika: $T_n(\mathbf{X}) \overset{H_0 \text{ esetén}}{\longrightarrow} \chi^2_{r-1-s}$ eloszlásban, ha $n \to \infty$.

Nagyon fontos: a próba aszimptotikus, tehát túl kicsi mintanagyság esetén nem hajtható végre a jelenlegi formájában. Nem egyértelmű, milyen határvonalat húzzunk meg. Hüvelykujjszabályként azt lehet mondani, hogy a minta elemszáma legyen legalább 50, az egyes cellákban pedig legyen legalább 5 legyen a gyakoriság. Általánosan korlátként lehet alkalmazni még a gyakoriságokra az $\sqrt[5]{n}$ számot. Amennyiben a cellákban túl alacsony a gyakoriságok száma, akkor az érintett osztályokat össze kell vonni másokkal.

b.) Diszkrét homogenitásvizsgálat

Feladat: van két **független** minta, mindkettő egy közös szempont szerint r osztály egyikébe sorolva. Azt kell eldönteni, hogy a két minta azonos eloszlásúnak tekinthető-e.

	Osztályok	1	2	...	r	Összesen
1. minta	Valószínűségek	p_1	p_2	...	p_r	1
	Gyakoriságok	N_1	N_2	...	N_r	n
2. minta	Valószínűségek	q_1	q_2	...	q_r	1
	Gyakoriságok	M_1	M_2	...	M_r	m

H_0 : a valószínűségek: $(p_1, \dots, p_r) = (q_1, \dots, q_r)$
H_1: nem ezek a valószínűségek

A próbastat.: $T_{n,m}(\mathbf{X}, \mathbf{Y}) = nm \sum_{i=1}^{r} \frac{\left(\frac{N_i}{n} - \frac{M_i}{m}\right)^2}{N_i + M_i} \overset{H_0 \text{ esetén}}{\longrightarrow} \chi^2_{r-1}$ elo.-ban, ha $n \to \infty$

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

c.) Függetlenségvizsgálat

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

$p_{i,j}$=P(egy megfigyelés az (i,j) osztályba kerül)
$N_{i,j}$=ennyi megfigyelés kerül az (i,j) osztályba

Itt formálisan a mintánk két dimenziós: a megfigyelések az $(X_1, Y_1)^T, \dots, (X_n, Y_n)^T$, ahol az X-ek r, az Y-ok pedig s különböző értéket vehetnek fel nemnulla valószínűséggel: $p_{i,j} = P(X_1 = x_i, Y_1 = y_j)$, ahol $i = 1, \dots, r$ és $j = 1, \dots, s$. Továbbá $N_{i,j} = \sum_{k=1}^{r} \sum_{l=1}^{s} P(X_k = x_j, Y_l = y_j)$.

A mintavétel eredménye:

		2. szempont					
		1	...	j	...	s	Összesen
1. szempont	1	N_{11}	...	N_{1j}	...	N_{1s}	$N_{1\bullet}$
	⋮	⋮		⋮		⋮	⋮
	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

ahol $N_{i\bullet} = \sum_{j=1}^{s} N_{ij}$ és $N_{\bullet j} = \sum_{i=1}^{r} N_{ij}$

H_0 : a szempontok függetlenek, azaz $p_{i,j} = p_{i\bullet} \cdot p_{\bullet j}$ $\forall i, j$-re
H_1: nem azok

Próbast.: $T_n(\mathbf{X}, \mathbf{Y}) = n \left(\sum_{i=1}^{r} \sum_{j=1}^{s} \frac{N_{i,j}^2}{N_{i\bullet} N_{\bullet j}} - 1 \right) \overset{H_0 \text{ esetén}}{\underset{n \to \infty}{\longrightarrow}} \chi^2_{(r-1)(s-1)}$ elo.-ban

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

Ha $r = s = 2$, akkor a próbastatisztika $T_n = n \cdot \frac{(N_{11}N_{22} - N_{12}N_{21})^2}{N_{1\bullet}N_{2\bullet}N_{\bullet 1}N_{\bullet 2}}$-re egyszerűsödik, az aszimptotikus eloszlás pedig 1 szabadságfokú χ^2.

4.) Nemparaméteres próbák folytonos esetben

a.) Folytonos illeszkedésvizsgálat – Kolmogorov-Szmirnov próba

Azt akarjuk ellenőrizni, hogy egy $X_1, \dots, X_n$ független, azonos eloszlású minta egy adott (fix paraméterű) folytonos eloszlásból származik-e. Tehát formálisan

$H_0 : F_{X_1}(x) = F(x) \quad \forall x \in \mathbb{R}$, ahol F egy adott eloszlás eloszlásfüggvénye
$H_1 : \exists x \in \mathbb{R} : F_{X_1}(x) \neq F(x)$

Próbastatisztika: $\sqrt{n} D_n(\mathbf{X}) = \sqrt{n} \cdot \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$

A próbastatisztika eloszlása H_0 esetén az ún. Kolmogorov-eloszláshoz tart ($n \to \infty$). Jelöljük K_α-val a Kolmogorov-eloszlás α-kvantilisét.

A kritikus tartomány: $\mathcal{X}_k = \{\mathbf{X} : \sqrt{n} D_n(\mathbf{X}) > K_{1-\alpha}\}$

Megj.: D_n kiszámításához elég csak a mintapontokban tekinteni az eltérést.
Megj.: azt is megtehetjük, hogy a mintából osztályközös gyakorisági sort hozunk létre – azaz mesterséges osztályozást készítünk –, majd χ^2-próbát hajtunk végre. Ezt az eljárást *diszkretizálás*nak hívjuk.

b.) Folytonos homogenitásvizsgálat

Van két független, azonos eloszlású mintánk: $X_1, \ldots, X_n$ és $Y_1, \ldots, Y_m$, mindkettő folytonos eloszlásból. Jelölje $F(x) = F_{X_1}(x)$ és $G(y) = F_{Y_1}(y)$. Az a célunk, hogy megállapítsuk, a két minta azonos eloszlású lehet-e, azaz az eloszlásfüggvényeik egyenlők-e. Tehát formálisan

$H_0 : F(x) = G(x) \quad \forall x \in \mathbb{R}$

$H_1 : F(x) \neq G(x)$ valamely $x \in \mathbb{R}$-re

Próbastatisztika: $\sqrt{\frac{nm}{n+m}} D_{n,m}(\mathbf{X}, \mathbf{Y}) = \sqrt{\frac{nm}{n+m}} \cdot \sup_{x \in \mathbb{R}} |F_n(x) - G_m(x)|$

A próbastatisztika eloszlása H_0 esetén a Kolmogorov-eloszláshoz tart $(n, m \to \infty)$.

A kritikus tartomány: $\mathcal{X}_k = \{(\mathbf{X}, \mathbf{Y}) : \sqrt{\frac{nm}{n+m}} D_{n,m}(\mathbf{X}, \mathbf{Y}) > K_{1-\alpha}\}$