Segédanyag a Matematikai statisztika tantárgyhoz

Leíró statisztika

Statisztikai **sokaság**: a megfigyelés tárgyát képező egyedek összessége, halmaza. Röviden sokaságnak hívjuk.
A sokaság egysége: a sokaság egy eleme.

Statisztikai **ismérv** (röv.: ismérv): a sokaság egyedeit jellemző tulajdonság.
Az ismérvek típusai:

- minőségi ismérv: az egyedek számszerűen nem mérhető tulajdonsága
- mennyiségi ismérv: az egyedek számszerűen mérhető tulajdonsága. Két fajtájukat különböztetjük meg:
  - diszkrét: véges vagy megszámlálhatóan sok értéket vehet fel
  - folytonos: egy adott intervallumon belül kontinuum számosságú értéket felvehet
- időbeli ismérv: az egységek időbeli elhelyezésére szolgáló rendezőelvek
- területi ismérv: az egységek térbeli elhelyezésére szolgáló rendezőelvek

Mérési szintek:

- Névleges (nominális) mérési skála: a számok csak ún. kódszámok, amik a sokaság egyedeinek azonosítására szolgálnak. Ezek között matematikai relációkat és műveleteket nincs értelme végezni. Pl. a hallgatók neme.
- Sorrendi (ordinális) skála: a sokaság egyedeinek valamely tulajdonság alapján sorba való rendezése. Pl. a hallgatók jegyei egy tárgyból.
- Intervallumskála (különbségi skála): a skálaértékek különbségei is valós információt adnak a sokaság egyedeiről. A skálán a nullpont meghatározása önkényes. Ilyen skálákhoz mértékegység is tartozik. Pl. hőmérséklet.
- Arányskála: a skálának van valódi nullpontja is. Minden matematikai művelet elvégezhető ezekkel a számokkal. Pl. a hallgatók magassága.

Tipikusan a minőségi ismérvek mérési szintje nominális, esetleg sorrendi skála; a mennyiségi ismérvek mérési szintje különbségi vagy arányskála; a területi ismérvek mérési szintje nominális skála; az időbeli ismérveké pedig különbségi skála.
Néha az intervallum- vagy arányskálán mérhető tulajdonságokat *metrikus ismérv*eknek nevezik.

Statisztikai tábla a statisztikai sorok összefüggő rendszere.
A statisztikai táblák fajtái:

- Egyszerű tábla: nincs benne összegző sor
- Csoportosító tábla: egyetlen összegző sort tartalmaz
- Kombinációs vagy *kontingenciatábla*: legalább két összegző sort tartalmaz

A statisztikai elemzések egyik legfontosabb eszközei a viszonyszámok. A **viszonyszám** két statisztikai adat hányadosa. Jelölések: $V = \frac{A}{B}$, ahol
V: viszonyszám; A: a viszonyítás tárgya; B: a viszonyítás alapja.

A viszonyszámok fajtái:

- Megoszlási: a sokaság egy részét a sokaság egészéhez viszonyítjuk.
- Koordinációs: a sokaság egy részének a sokaság egy másik részéhez való viszonyítása.
- Dinamikus: két időpont vagy időszak adatának hányadosa.
- Intenzitási: különböző fajta adatok viszonyítása egymáshoz; gyakran a mértékegységük is eltérő.

Ha egy teljes sokaságra és annak m részére rendelkezésre áll a viszonyszám alapja és részei, akkor a viszonyszámokat ki tudjuk számolni a teljes sokaságra (jel. $\overline{V}$, ezt *összetett viszonyszám*nak hívják) és annak részeire is (jel. $V_1, \dots, V_m$). Ekkor a teljes sokaságra számolt viszonyszám kiszámítási lehetőségei:

$$\overline{V} = \frac{\sum_{i=1}^{m} A_i}{\sum_{i=1}^{m} B_i} = \underbrace{\frac{\sum_{i=1}^{m} B_i V_i}{\sum_{i=1}^{m} B_i}}_{\substack{\text{súlyozott} \\ \text{számtani átlag}}} = \underbrace{\frac{\sum_{i=1}^{m} A_i}{\sum_{i=1}^{m} \frac{A_i}{V_i}}}_{\substack{\text{súlyozott} \\ \text{harmonikus átlag}}}$$

A leíró statisztikai szakirodalomban az i indexeket – pongyola módon – le szokták hagyni:

$$\overline{V} = \frac{\sum A}{\sum B} = \frac{\sum BV}{\sum B} = \frac{\sum A}{\sum \frac{A}{V}}$$

Definíció. **z-kvantilis:** $q(z) = q_z = \inf\{x : F(x) \geq z\}$, és amennyiben F invertálható, akkor $q_z = F^{-1}(z)$-re egyszerűsödik $(0 < z < 1)$
Fontos speciális kvantilisek: kvartilisek:

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

Definíció. **Módusz:** abszolút folytonos eloszlás esetén a sűrűségfüggvény maximumhelye(i), diszkrét eloszlás esetén pedig az eloszlás maximumhelye(i). Tehát

- Mo= $\underset{x\in\mathbb{R}}{\operatorname{argmax}}\ f(x)$, ha X abszolút folytonos;
- Mo= $\underset{x_1,x_2,\ldots}{\operatorname{argmax}}\ P(X=x_i)$, ha X diszkrét.

Nem biztos, hogy létezik, és ha létezik, akkor se biztos, hogy egyértelmű.

Definíció. **Ferdeség (skewness):** $\operatorname{skew}(X) = \frac{E(X-EX)^3}{(DX)^3}$

Értelmezése:
- skew(X)=0 $\Rightarrow$ az eloszlás szimmetrikus
- skew(X)¿0 $\Rightarrow$ az eloszlás balra ferdült
- skew(X)¡0 $\Rightarrow$ az eloszlás jobbra ferdült

Definíció. **Csúcsosság (kurtosis):** $\operatorname{kurt}(X) = \frac{E(X-EX)^4}{(DX)^4} - 3$

Értelmezés:
- kurt(X)=0 $\Rightarrow$ az eloszlás csúcsossága a standard normáliséval megegyező
- kurt(X)¡0 $\Rightarrow$ az eloszlás laposabb a st. norm.-nál
- kurt(X)¿0 $\Rightarrow$ az eloszlás csúcsosabb a st. norm.-nál

Minta: $X_1, \ldots, X_n$ valószínűségi változó sorozat, jel. $\mathbf{X} = (X_1, \ldots, X_n)^T$
A továbbiakban feltesszük, hogy függetlenek és azonos eloszlásúak – ezt röviden *i.i.d. mintá*nak hívjuk (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: $T : \mathbf{X} \mapsto ...$
Becslés: a minta eloszlásának ismeretlen paraméterét közelíti a minta segítségével.
Megj.: Minden becslés statisztika.
Néhány lényeges statisztika:

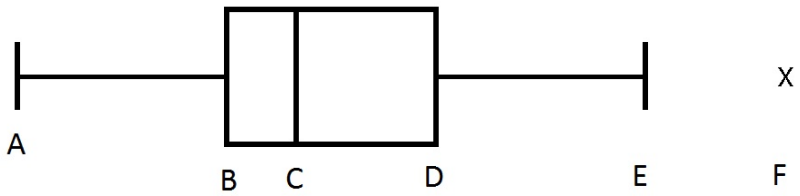
- *Rendezett minta*: $X_1^* \le ... \le X_n^*$ nem csökkenő sorrendbe tesszük a mintaelemeket
- *Terjedelem*: $R = X_n^* - X_1^*$ (R=range)
- *Mintaátlag*: $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
- *Tapasztalati szórás*: $S_n = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n}}$
  Értelmezése: 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*: $V = \frac{S_n}{\overline{X}}$
  Értelmezése: az átlagtól való átlagos eltérés százalékban
  Megj.: relatív szórásnak is hívják
- *Tapasztalati eloszlásfüggvény*: $F_n(x) = \frac{\sum_{i=1}^{n} I(X_i<x)}{n}$
  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
- *Tapasztalati z-kvantilis*: Realizált mintából sokféleképpen számolható, interpolációs módszer:

  1.) Sorszám megállapítása: $(n+1)z = e + t$ (e: egészrész, t: törtrész)
  2.) $q_z = x_e^* + t(x_{e+1}^* - x_e^*)$

  Értelmezése: a mintaelemek z-ed része legfeljebb a q_z értéket veszi fel, $(1-z)$-ed része pedig legalább q_z.
- *Interkvartilis terjedelem*: $IQR = Q_3 - Q_1$
- *Tapasztalati módusz*: a legtöbbször előforduló érték.
  Értelmezése: a minta tipikus, leggyakrabban előforduló értéke.
- *Tapasztalati ferdeség*: $\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^3}{S_n^3}$
- *Tapasztalati csúcsosság*: $\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^4}{S_n^4} - 3$

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

Boxplot ábra: (ez fekvő, de lehet álló is)
ahol a betűk a következő értékeket jelentik:



- $A = \max\{x_1^*, Q_1 - 1,5 \cdot IQR\}$;
- $B = Q_1$;
- $C = Me$;
- $D = Q_3$;
- $E = \min\{x_n^*, Q_3 + 1,5 \cdot IQR\}$;
- F: kieső értékek, azokat tüntetjük fel pontokként, amik A-n vagy E-n kívülre esnek.

Az adatelemzés lépései:

- Adathibák keresése, irreális adatok, értékek törlése; esetleg korrigálása
- Alkalmas osztályközös gyakorisági sor készítése
- Középértékek kiszámítása
  - Átlag (számtani vagy mértani – amelyiknek értelme van)
  - Helyzeti középértékek: módusz és medián
- Szóródási mutatók kiszámítása
  - Terjedelem és interkvartilis terjedelem
  - Szórás és relatív szórás
- Alakmutatók kiszámítása
  - Ferdeség
  - Csúcsosság
- Ábrák készítése:
  - Sűrűséghisztogram
  - Boxplot ábra

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)	$P(X=k)=p(1-p)^{k-1}$ $k=1,2,\dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Hipergeometriai	Hipgeo(N,M,n)	$P(X=k)=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$ $k=0,1,\dots,n$	$n\frac{M}{N}$	$n\frac{M}{N}\left(1-\frac{M}{N}\right)\left(1-\frac{n-1}{N-1}\right)$
Binomiális	Bin(n,p)	$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ $k=0,1,\dots,n$	np	$np(1-p)$
Negatív binomiális	NegBin(n,p)	$P(X=k)=\binom{k-1}{n-1}p^n(1-p)^{k-n}$ $k=n,n+1,\dots$	$\frac{n}{p}$	$\frac{n(1-p)}{p^2}$
Poisson	Poi(λ)	$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$ $k=0,1,\dots$	λ	λ

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 norm.	N($0,1^2$)	$\Phi(x)=\dots$	$\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\quad x\in\mathbb{R}$	0	1
Normális	N(m,σ^2)	$\dots$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}\quad x\in\mathbb{R}$	m	σ^2

További nevezetes abszolút folytonos eloszlások:

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

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

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	$\Gamma(\alpha,\lambda) \quad \alpha,\lambda > 0$	...	$\begin{cases} \frac{1}{\Gamma(\alpha)} \lambda^\alpha e^{-\lambda x} x^{\alpha-1} & \text{ha } x \geq 0 \\ 0 & \text{ha } x < 0 \end{cases}$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
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	$\mathrm{LN}(m,\sigma^2)$ $m \in \mathbb{R}, \sigma > 0$	...	$\begin{cases} \frac{1}{x\sqrt{2\pi}\sigma} e^{-\frac{(\log x - m)^2}{2\sigma^2}} & \text{ha } x \geq 0 \\ 0 & \text{ha } x < 0 \end{cases}$	$e^{m+\sigma^2/2}$	$(e^{\sigma^2-1})e^{2m+\sigma^2}$

Matematikai statisztika – becsléselmélet

Most belekezdünk a matematikai statisztikába, a korábbi minta fogalma egy fokkal absztraktabb formában fog visszaköszönni.

Definíció. **Statisztikai mező.** $(\Omega, \mathcal{A}, \mathcal{P})$ hármas, ahol $\mathcal{P}$ pedig eloszlások egy családja és minden $P \in \mathcal{P}$-re $(\Omega, \mathcal{A}, P)$ valószínűségi mező.
$\mathcal{P}$-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.** $\mathbf{X} : (\Omega, \mathcal{A}) \longrightarrow \mathcal{X}$ leképezés, ahol $\mathcal{X}$ neve: **mintatér**.

Feladat: annak a meghatározása, hogy a $\mathcal{P}$ eloszláscsalád melyik tagja írja le legjobban a valóságot, a vizsgált jelenséget. Ennek érdekében veszünk mintát. Erőfeszítéseink jelentős része arra fog irányulni, hogy a valamilyen szempontból "legjobb" P-t vagy paraméteres esetben ezzel ekvivalens módon, a "legjobb" ϑ paramétert megtaláljuk.
Jelölés. A továbbiakban a valószínűség, sűrűségfüggvény, várható érték és szórás(mátrix) alsó indexben lévő ϑ arra fog utalni, hogy egy paraméteres statisztikai mező van a feladat hátterében és ϑ-val jelöljük az ismeretlen, de érdeklődésünk középpontjában lévő paramétert.

Definíció. **Likelihood függvény:** Legyen $\mathbf{X} = (X_1, \ldots, X_n)$ 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. **Log-likelihood függvény:** $l(\vartheta; \mathbf{x}) = \log L(\vartheta; \mathbf{x})$

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

Definíció. **Torzítatlan becslés.** $\mathrm{T}(\mathbf{X})$ statisztika 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 $D^2_\vartheta(T_1(\mathbf{X})) \leq D^2_\vartheta(T_2(\mathbf{X}))$ minden $\vartheta \in \Theta$ esetén.

Definíció. **Hatásos becslés.** A $T(\mathbf{X})$ torzítatlan 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 $P_\vartheta(T_1(\mathbf{X}) = T_2(\mathbf{X})) = 1 \quad \forall \vartheta \in \Theta$ esetén.

Megjegyzés. Egy becslésről nem egyszerű belátni, hogy hatásos. Hatásos becslés keresésének alapja a Blackwell-Rao tétel.

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

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

Állítás.

- A tapasztalati eloszlásfüggvény torzítatlan és erősen konzisztens becslése az eloszlásfüggvénynek.
- A mintaátlag torzítatlan és erősen konzisztens becslése a várható értéknek.
- A tapasztalati szórásnégyzet *torzított*, de aszimptotikusan torzítatlan és erősen konzisztens becslése a szórásnégyzetnek.
- A tapasztalati szórásnégyzet torzítatlan és erősen konzisztens becslése a korrigált szórásnégyzetnek.

Paraméterbecslési módszerek

- **Maximum likelihood módszer (ML-módszer):** Azt a paraméterértéket keressük, ahol a likelihood fv. a legnagyobb értéket veszi fel: $\max_{\vartheta} L(\vartheta; \mathbf{x})$
  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.

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

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

Definíció. **Elégséges statisztika.**
Legyen $(\Omega, \mathcal{A}, \mathcal{P})$ statisztikai mező, $\mathbf{X}$ minta, $B \in \mathcal{A}$. A T statisztikát **elégséges statisztiká**nak nevezzük, ha a $P_\vartheta(\mathbf{X} \in B | T(\mathbf{X}))$ feltételes eloszlásnak létezik ϑ-tól nem függő változata.
Megjegyzés. Ez egy elég absztrakt fogalom. Elégséges statisztikát a Neyman-féle faktorizációs tétel segítségével (kicsit lejjebb) tudunk keresni és arra lesz jó, hogy segítségével bizonyos szempontból optimális becslést találjunk.
Megjegyzés. az elégséges statisztika minden lényeges információt tartalmaz az ismeretlen ϑ paraméterre vonatkozóan.

Tétel. **Neyman-féle faktorizációs tétel.**
"Szép" statisztikai mezőn a T statisztika akkor és csak akkor elégséges, ha léteznek olyan g_ϑ nemnegatív és h 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.

Állítás. A $T(\mathbf{X}) = \mathbf{X}^*$ rendezett minta elégséges statisztika.
Ebben a részben tegyük fel, hogy a paramétertér 1 dimenziós.

Definíció. **Fisher-információ.**
Tegyük fel, hogy a log-likelihood függvény ϑ szerint deriválható. Ekkor az $\mathbf{X}$ n elemű mintában lévő Fisher-információ: $I_{\mathbf{X}}(\vartheta) \equiv I_n(\vartheta) = E_\vartheta\left([\partial_\vartheta l(\vartheta; \mathbf{X})]^2\right)$.

Megj.: $I_{\mathbf{X}}(\vartheta)$ azt az (absztrakt) információmennyiséget méri, amelyet az $\mathbf{X}$ minta a paraméterre vonatkozóan magában hordoz.
A Fisher-információ kiszámítása bizonyos, úgynevezett regularitási feltételek esetén egyszerűbbé válik.

Definíció. **1. regularitási feltétel.** $E_\vartheta(\partial_\vartheta l(\vartheta, \mathbf{X})) = \mathbf{0}$

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

Állítás. Ha teljesül az 1. regularitási feltétel, akkor a Fisher-információt kiszámolhatjuk az alábbi módon: $I_n(\vartheta) = n \cdot I_1(\vartheta) = n \cdot D^2_\vartheta \partial_\vartheta l(\vartheta, X_1)$.

Tétel. **Cramér-Rao egyenlőtlenség.**
Tegyük fel, hogy $T(\mathbf{X})$ statisztika torzítatlan becslése $g(\vartheta)$-nak és teljesül az 1. regularitási feltétel. Ekkor minden $\vartheta \in \Theta$-ra $D^2_\vartheta(T(\mathbf{X})) \geq \underbrace{\frac{(g'(\vartheta))^2}{I_n(\vartheta)}}_{\text{információs határ}}$.

Megjegyzés. Ha minden ϑ-ra egyenlőség teljesül a Cramér-Rao egyenlőtlenségben, akkor T hatásos becslés. Ennek az egyenlőtlenségnek a vizsgálata tehát lehetőséget ad arra, hogy blackwellizálás nélkül hatásos becslést találjunk.

Megjegyzés. Előfordulhat, hogy a statisztika szórásnégyzete nagyobb az információs határnál, viszont a statisztika hatásos. Példa erre i.i.d. exponenciális mintánál az $\frac{\frac{n-1}{n}}{\sum\limits_{i=1}^{n} X_i}$ statisztika.

Definíció. **χ^2-eloszlás:** Az X val. változó n szabadságfokú χ^2-eloszlású (jel.: $X \sim \chi^2_n$), ha $X = U_1^2 + \ldots + U_n^2$, ahol $U_i \sim N(0,1)$ $\forall i$-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.

Definíció. **F-eloszlás:** Az X valószínűségi változó m és n szabadságfokú F-eloszlást követ (jel.: $X \sim F_{m,n}$), ha $X = \frac{\frac{Y_m}{m}}{\frac{Z_n}{n}}$, ahol $Y_m \sim \chi^2_m$ és $Z_n \sim \chi^2_n$ függetlenek egymástól.
Mostantól α egy 0-hoz közeli pozitív szám lesz (például $0,05 = 5\%$), és vezessük be a következő jelöléseket az eloszlások kvantiliseire:

- 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
- $F^{\alpha}_{m,n}$: m,n szabadságfokú F-eloszlás α-kvantilise

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

Állítás. Legyen $X_1, \ldots, X_n \sim N(m, \sigma^2)$ i.i.d. minta. Ekkor

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

Intervallumbecslés az ismeretlen ϑ paraméterre a paraméter ML-becslésének aszimptotikája alapján n elemű mintából: $\hat{\vartheta} \pm u_{\frac{\alpha}{2}} \cdot \frac{1}{\sqrt{I_n(\hat{\vartheta})}}$

Hipotézisvizsgálat

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:
$H_0: \ \vartheta \in \Theta_0$ $\rightsquigarrow$ nullhipotézis
$\underline{H_1: \ \vartheta \in \Theta_1}$ $\rightsquigarrow$ ellenhipotézis vagy alternatív hipotézis

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: $\mathcal{X} \to \mathbb{R}$, $\mathcal{X}_k = \{\mathbf{x} \in \mathcal{X} : \mathrm{T}(\mathbf{x}) > c\}$ c neve: kritikus érték
$\mathcal{X}_e = \{\mathbf{x} \in \mathcal{X} : \mathrm{T}(\mathbf{x}) \leq c\}$

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

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.

A hipotézisvizsgálat menete próbastatisztika és kritikus érték segítségével:

1. A terjedelem (α) lefixálása, ami jellemzően 1% és 10% közötti, tipikusan 5%
   Megbízhatóság= $1-\alpha$, általában %-osan írjuk
2. Nullhipotézis (H_0) felírása – sokévi, megszokott, elvárt értékeknek megfelelő paramétertartomány
3. Alternatív hipotézis (H_1) felírása – a minta alapján bennünket érdeklő kérdésnek megfelelő paramétertartomány

4. A probléma megoldására alkalmas próba vagy próbák kiválasztása – feltételek ellenőrzése
5. Próbastatisztika kiszámítása
6. Kritikus érték kiszámítása, kritikus tartomány ($\mathcal{X}_k$) megállapítása
7. Döntés:
   (a) $\mathbf{x} \in \mathcal{X}_k \rightsquigarrow$ **erős döntés**, H_1-et elfogadjuk, H_0-t elvetjük/elutasítjuk
   (b) $\mathbf{x} \in \mathcal{X}_e \rightsquigarrow$ **gyenge döntés**, H_1-et elutasítjuk, H_0-t nem tudjuk elutasítani
8. Szöveges értelmezés: α terjedelem/elsőfajú hiba valószínűsége mellett azt állíthatjuk/nem állíthatjuk, hogy ...

A hipotézisvizsgálat menete p-érték segítségével:

1. A terjedelem (α) lefixálása
2. Nullhipotézis (H_0) felírása
3. Alternatív hipotézis (H_1) felírása
4. A probléma megoldására alkalmas próba vagy próbák kiválasztása
5. p-érték megállapítása, rendszerint számítógép segítségével
6. Döntés:
   (a) p-érték $< \alpha \quad \Leftrightarrow \quad \mathbf{x} \in \mathcal{X}_k \quad \Leftrightarrow \quad H_1$-et elfogadjuk
   (b) p-érték $> \alpha \quad \Leftrightarrow \quad \mathbf{x} \in \mathcal{X}_e \quad \Leftrightarrow \quad H_0$-t nem tudjuk elutasítani
7. Szöveges értelmezés

Most néhány nevezetes próbát mutatunk be a normális eloszlás várható értékére, illetve szórására. Az α végig a próba terjedelmét jelöli, ami előre adott.

I. Próbák normális eloszlás várható értékére

1.) Egymintás próbák

a.) Egymintás u-próba

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

$X_1, \dots, X_n \sim \mathrm{N}(m, \sigma^2)$, 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	a.) kétmintás u-próba		egymintás u-próba a különbségekre
σ_1 és σ_2 ismeretlen	előzetes F-próba		
	$\sigma_1 = \sigma_2$	$\sigma_1 \neq \sigma_2$	egymintás t-próba a különbségekre
	b.) kétmintás t-próba	c.) Welch-próba	

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

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

c.) 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: $t' = \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{(s_1^*)^2}{n}+\frac{(s_2^*)^2}{m}}} \overset{H_0 \text{ esetén}}{\sim} t_f$, 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 $s_1^* > s_2^*$

II. Próbák normális eloszlás szórására

1.) **Egymintás próba:** χ^2-próba

$X_1, \ldots, X_n \sim N(m, \sigma^2)$, ahol m és σ ismeretlen paraméterek

$H_0 : \sigma = \sigma_0$ és $H_1 : \sigma \neq \sigma_0$

A próbastatisztika: $h = \frac{(n-1)(S_n^*)^2}{\sigma_0^2} \overset{H_0 \text{ esetén}}{\sim} \chi^2_{n-1}$

Kritikus tartomány: $\mathcal{X}_k = \left\{ \mathbf{x} : h < \chi^2_{n-1,\alpha/2} \text{ vagy } h > \chi^2_{n-1,1-\alpha/2} \right\}$

Az ellenhipotézis lehet egyoldali is, ilyenkor a kritikus tartomány értelemszerűen módosul.

2.) **Kétmintás próba:** F-próba

$X_1, \ldots, X_n \sim \mathrm{N}(m_1, \sigma_1^2)$ $Y_1, \ldots, Y_m \sim \mathrm{N}(m_2, \sigma_2^2)$ $m_1, m_2, \sigma_1, \sigma_2$ paraméterek

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

A próbastatisztika: $F = \frac{(S_1^*)^2}{(S_2^*)^2} \overset{H_0 \text{ esetén}}{\sim} F_{n-1,m-1}$

χ^2-próbák

a.) Diszkrét illeszkedésvizsgálat

Feladat: adott egy $\mathbf{X} = (X_1, \ldots, X_n)$ 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: $T_n = \sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i} \overset{H_0 \text{ esetén}}{\longrightarrow} \chi^2_{r-1}$ 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 $n \to \infty$.

Nagyon fontos: a próba csak akkor hajtható végre, amennyiben az egyes osztályokban elegendő számú gyakoriság szerepel. Nem egyértelmű, milyen határvonalat húzzunk meg. Hüvelykujjszabályként azt lehet mondani, hogy legalább 4-6 gyakoriság szerepeljen a cellákban és np_i legalább 4 legyen minden osztályra. Amennyiben kevés gyakoriság van a cellákban, akkor az érintett osztályokat össze kell vonni.

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: $(p_1, \ldots, p_r) = (q_1, \ldots, q_r)$

H_1: nem ezek a valószínűségek

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

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

c.) Függetlenségvizsgálat

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

$\mathrm{p}_{i,j}$=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. 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

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

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.

Folytonos illeszkedésvizsgálat

Azt akarjuk ellenőrizni, hogy egy $X_1, \ldots, X_n$ 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
$H_1 : \exists x \in \mathbb{R} : F_{X_1}(x) \neq F(x)$

Kolmogorov-Szmirnov próba
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 ($n \to \infty$), melynek eloszlásfüggvénye $1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2}$. 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.

Nemparaméteres próbák az u-próbák/t-próbák helyett

Ezek a próbák az egymintás és kétmintás u-próbák/t-próbák helyett hajtandók végre, amennyiben nem teljesül, hogy a minta normális eloszlású (például vannak benne kiugró értékek). Az egymintás próba erre a célra az előjel próba és a Wilcoxon-próba, a kétmintás esetben pedig független minták esetén a Mann-Whitney próba. Kétmintás esetben amennyiben összefüggők a minták, akkor a különbségekre Wilcoxon-próbát vagy előjel próbát hajtunk végre.

1.) Egymintás próbák

Adott egy $X_1, \ldots, X_n$ folytonos eloszlásból származó minta. A nullhipotézis azt állítja, hogy a medián egy előre megadott m_0 számmal egyenlő-e. Az ellenhipotézis egy- és kétoldali is lehet.
A nullhipotézis: $H_0 : m = m_0$, ahol m a mediánt jelöli. Másképp: $H_0 : P(X - m_0 > 0) = 1/2$

a.) Előjel próba

A próbastatisztika: $T = \sum_{i=1}^{n} I(X_i > m_0)$ $\rightsquigarrow$ megszámoljuk, hány mintaelem nagyobb m_0-nál A próbastatisztika H_0 teljesülése esetén $Bin\left(n, \frac{1}{2}\right)$ eloszlású, ezáltal a p-érték

- kétoldali ellenhipotézis esetén: $\frac{2}{2^n}\sum_{i=0}^{min(T,n-T)}\binom{n}{i}$
- baloldali ellenhipotézis esetén: $\frac{1}{2^n}\sum_{i=0}^{T}\binom{n}{i}$
- jobboldali ellenhipotézis esetén: $\frac{1}{2^n}\sum_{i=0}^{n-T}\binom{n}{i}$

b.) Wilcoxon-próba

A próba végrehajtásához legyen $Y_i := |X_i - m_0|, \quad i = 1, \ldots, n$. Jelölje R_i az i-edik mintaelem rangját: $R_i = \sum_{j=1}^{n} I(Y_i \geq Y_j)$.
A próbastatisztika azon rangok összege, amikre $X_i - m_0$ pozitív: $T^+ = \sum_{i:\ X_i > m_0} R_i$
A kritikus értékek meghatározása viszonylag macerás. Kis mintákra speciális táblázatokból kell kinézni az értékeket, nagy mintákra pedig a nullhipotézis esetén a lenormált próbastatisztika közel standard normális eloszlású, így ennek a kvantiliseit lehet használni.

2.) Kétmintás próbák Adott egy $X_1, \ldots, X_n$ minta és egy tőle független $Y_1, \ldots, Y_m$ minta, mindkettő folytonos eloszlásból. A nullhipotézis azt állítja, hogy a két eloszlás mediánja megegyezik-e. Az ellenhipotézis egy- és kétoldali is lehet.
A nullhipotézis: $H_0 : P(X > Y) = P(X < Y)$.

a.) Mann-Whitney próba

A próbastatisztika: $W = \sum_{i=1}^{n}\sum_{j=1}^{m} I(X_i > Y_j)$.
A kritikus értékek meghatározása hasonlóan megy, mint a Wilcoxon-próbánál. Kis mintákra speciális táblázatokból kell kinézni az értékeket, nagy mintákra pedig a próbastatisztikát le kell normálni és használhatjuk a standard normális eloszlás kvantiliseit.