Összefoglaló

Matematikai statisztika

Bevezetés

Statisztikai sokaság, populáció

Definíció: A vizsgálat tárgyát képező nagyszámú, de *véges* elemszámú egyedek halmaza. A halmaz egészének kevés adattal történő tömör jellemzése, és a populáció egyedeinek leírására bevezetett változók közötti kapcsolatok leírása a célunk. Arra nincs lehetőség (erőforrás), hogy a populáció minden egyes eleméről adatokat szerezzünk be.

Példa:

- Magyarország állampolgárai
- Egy egyetemi kar hallgatói
- Az érvényes forgalmival rendelkező autók halmaza
- Egy adott termék vásárlóinak halmaza
- Egy TV csatorna nézőinek halmaza

A statisztikai elemzés tárgya lehet egy véletlen kísérlet is, ami időben változatlan körülmények között elvileg akárhányszor lejátszódhat.

K	véletlen kísérlet

Példa:

- A lottóhúzás
- Egy szerver működése
- Budapest januári átlaghőmérséklete
- Egy gyümölcsös terméshozama
- Egy új gyógyszer hatása
- Egy reklámkampány hatásossága
- Egy populáció egyedének véletlen kiválasztása

Minta realizáltja

Definíció: A populáció egy kis elemszámú részhalmazára vonatkozó megfigyelések adatai. A minta úgy kell, hogy tükrözze a populáció tulajdonságait, ahogy a cseppben látjuk a tengert. Azaz a minta reprezentatív kell, hogy legyen.

Példa:

- Egy felmérésbe bevont magyar állampolgárok halmaza
- Egy adott előadásra belátogatott hallgatók halmaza
- Adott biztosítóval szerződött autók halmaza
- Egy adott napon megkérdezett vásárlók halmaza
- Egy nézettségi felmérésbe bevont TV nézők halmaza
- Budapest januári középhőmérsékleteinek adatai

Mintavételezési eljárások

Szabályok:

- A populáció minden egyes elemének ugyanakkora esélyt kell biztosítani a mintába kerüléshez.
- A minta elemszámának elég nagynak kell lennie ahhoz, hogy a következtetéseink átvihetők lehessenek a populációra is.

Típusai:

- **Rétegzett mintavételezés**: A populációt adott szempontok szerint csoportokba osztjuk, és a csoportok arányait a mintában is megtartjuk
- **Véletlen mintavételezés**: A mintába kerülő egyedeket sorsolással választjuk ki.
- **Cenzus**: népszámlálás

Alapfogalmak

Eset: A minta egy eleme, az adatmátrix egy sora.

Mintaelemszám: Az adott minta elemeinek száma. Egy adatmátrix sorainak száma.

Adatmátrix: *n* db eset és *p* db változó adatainak mátrixba rendezett alakzata

Változó: A populáció egy mérhető jellemzője. Az adatmátrix egy oszlopa.

A minta realizáció adataiból adott képlettel számolt adat a statisztika számított értéke: átlag, standard szórás, medián, kvartilis, ferdeség, lapultság, módusz, gyakoriság, próbastatisztikák, stb.

K	a véletlen kísérlet
Ω	a lehetséges kimenetelek halmaza
A	a megfigyelhető események halmaza
P	a lehetséges valószínűségi mértékek halmaza
ϑ	paramétervektor

Az elemzésünk célja, hogy a P halmazból kiválasszuk a tényleges valószínűséget! Legalább is egy jó helyettesítő egyedet.

Statisztikai minta

Definíció: Az X valószínűségi változóval azonos eloszlású, egymással teljesen független $X_1, X_2,\ldots, X_n$ valószínűségi változók együttesét *statisztikai mintának* nevezzük.

X	eloszlásfüggvény, a minta eloszlásfüggvénye
n	mintaelemszám
X_i	a minta i-edik eleme
ω	realizálódott kimenet

Egy mintavételezéskor tulajdonképpen megfigyeljük a K véletlen kísérletet, azaz megállapítjuk melyik $\omega \in \Omega$ kimenetele realizálódott.

Az $X_1(\omega) = x_1$, $X_2(\omega) = x_2,\ldots, X_n(\omega) = x_n$ szám n-est nevezzük a minta realizációjának.

Statisztika

Definíció: Legyen t_n egy n-változós valós függvény. Akkor a statisztikai minta $T_n=t_n(X_1,X_2,\ldots,X_n)$ függvényét nevezzük *statisztikának*.

A statisztika egy valószínűségi változó, aminek eloszlásfüggvényét a minta eloszlásfüggvényéből lehet kiszámolni.

A $T_n = t_n(X_1, X_2,\ldots,X_n)$ szám (amikor az argumentumba mintarealizáció értékeit helyettesítjük,a statisztika számolt értéke.

Adatcentrum statisztikák

átlag – vagy empirikus közép	$\bar{X}_n$	$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
medián	m_n	$m_n = \begin{cases} x^*_{\frac{n+1}{2}}, ha\ n\ páratlan \\ \frac{x^*_{\frac{n}{2}} + x^*_{\frac{n}{2}+1}}{2}, ha\ n\ páros \end{cases}$
módusz	leggyakrabban előforduló érték	

Szóródást jellemző statisztikák

${s_n}^2$	${s_n}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$	empirikus szórásnégyzet
${s_n^*}^2$	${s_n^*}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$	korrigált empirikus szórásnégyzet
s_X^*	$s_X^* = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$	korrigált empirikus szórás

Empirikus eloszlásfüggvény

Az empirikus eloszlásfüggvény minden *x* helyen egy lépcsős eloszlásfüggvény lesz. Ugyanakkor az eloszlásfüggvény a statisztikai minta függvénye is, azaz minden *x* helyen valószínűségi változó lesz

Glivenko-Cantelli-tétel: Az empirikus eloszlásfüggvény 1 valószínűséggel, egyenletesen konvergál az eloszlásfüggvényhez.

Paraméterbecslés

Paraméter

Tegyük fel, hogy a minta eloszlásfüggvénye képletét egy ϑ paraméter konkretizálja. Ha ismerjük az értékét, meg tudjuk pontosan adni az eloszlásfüggvényt:

$$F = \{F_x(t, \vartheta) : \vartheta \in \mathbf{\Theta}\}$$

Célunk: Egy adott statisztikai minta segítségével a υ paraméter megbecslése.

Paraméter becslése

A ϑ paraméter becsléséhez valamilyen alkalmas T_n statisztikát használunk: $T_n \approx \vartheta$. Egy ismeretlen számot (a ϑ -át) egy valószínűségi változóval becsüljük! **Mikor jó egy ilyen becslés?**

Torzítatlanság

Egy valószínűségi változó az összes szám közül éppen a várható értéke körül ingadozik a legkisebb mértékben. A torzítatlanság azt jelenti, hogy a becslő statisztika éppen a becsülendő paraméterérték körül fogja felvenni az értékeit.

A T_n statisztika a ϑ paraméter torzítatlan becslése, ha $\mathbf{E}T_n = \vartheta$

Aszimptotikus torzítatlanság

Ha a torzítatlansági feltétel csak $n \to \infty$ esetben igaz:

$$\lim_{n\to\infty} ET_n = \vartheta$$

A torzítatlanságból nyilvánvalóan következik az aszimptotikus torzítatlanság, tehát az utóbbi gyengébb feltétel.

Konzisztencia

Ha minta elemszám növekedtével növekszik a becslés pontosságának valószínűsége, konzisztens becslésről beszélünk. Minden $\varepsilon < 0$ esetén teljesül, hogy:

$$\lim_{n\to\infty} P(|T_n - \vartheta| > \varepsilon) = 0$$

Erős konzisztencia

Azok a torzítatlan becslések, melyeknél a variancia a minta elemszám növekedtével 0-hoz tart:

$$ET_n = \vartheta \text{ és } \lim_{n\to\infty} D^2 T_n = 0$$

Csak a konstansnak lehet 0 a varianciája. Tehát, ha n elég nagy, a becslés gyakorlatilag a paramétert adja! A Csebisev – egyenlőtlenségből következik, hogy az erősen konzisztens statisztikai becslések egyben konzisztensek is lesznek. A megfordítás általában nem igaz!

Hatásosság

Két torzítatlan becslés közül nyilván a kisebb varianciájú a jobb, hiszen kisebb mértékben ingadozik a paraméter körül! Azaz, a V_n statisztika *hatásosabb* W_n-nél, ha:

$$EW_n = EW_n = \vartheta \text{ és } D^2 V_n \leq D^2 W_n$$

Egy torzítatlan becslés akkor lesz *hatásos*, ha varianciája minden más torzítatlan becslés varianciájánál kisebb! Csak egyetlen hatásos becslés van! (Ezt kell megkeresni egy adott paraméter-becslési problémához!)

Becsülendő paraméterek és értékelése

Várható érték

Becsülendő paraméter most az X várható értéke.

$$EX = \vartheta$$

Az átlagstatisztika torzítatlan.

Ha még azt is tudjuk, hogy $D^2 X < \infty$ akkor az átlag erősen konzisztens is.

A lineáris becslések között az átlag a hatásos.

Variancia

Becsülendő paraméter most az X varianciája:

$$D^2 X = \vartheta$$

Az empirikus szórásnégyzet statisztika a ϑ szórásnégyzet aszimptotikusan torzítatlan becslése, a korrigált empirikus szórásnégyzet pedig a ϑ szórásnégyzet torzítatlan becslése.

Összefoglalva:

- az **átlagstatisztika** a minta várható értékének –mint paraméternek- **torzítatlan** becslése. Ha a mintának létezik szórása, akkor ez a becslés **erősen konzisztens** is.
- A minta **tapasztalati szórásnégyzete** a minta varianciájának –mint paraméternek- **aszimptotikusan torzítatlan** becslése. Ha a mintának létezik negyedik momentuma, akkor a becslés **konzisztens** is.
- A minta **korrigált empirikus szórásnégyzet** statisztika a minta varianciájának **torzítatlan** becslése. Ha a minta negyedik momentuma létezik, akkor **erősen konzisztens** becslése.

Hipotézisvizsgálat

Döntési eljárást dolgozunk ki annak eldöntésére, hogy a nullhipotézis igaz-e. Ha úgy kell döntenünk, hogy a nullhipotézis nem igaz, automatikusan az alternatív hipotézist fogjuk elfogadni. A döntésünkhöz szignifikancia szintet fogunk rendelni, amivel jellemezzük, hogy a nullhipotézisünk melletti döntés milyen erős.

- Null hipotézis: H_0
- Alternatív hipotézis: H_1

Döntés - Valóság	H_0 igaz	H_1 igaz
H_0 elfogadjuk	Helyes döntés	Másodfajú hiba
H_1-t elfogadjuk	Elsőfajú hiba	Helyes döntés

Elsőfajú hiba: Akkor követjük el, ha igaz a nullhipotézis, de a mintrealizáció mégis a kritikus tartományba esik, és a döntésünk elutasító! Az elsőfajú hibavalószínűség ε, amit mi állítunk be!

Másodfajú hiba: Akkor követjük el, ha elfogadjuk a nullhipotézist, holott valójában nem igaz. Értéke nehezebben állapítható meg.

Paraméteres próbák

A próbákban az a közös, hogy az elemzett minta eloszlása **normálist** követ. A nullhipotézist éppen a normális eloszlás paramétereivel kapcsolatosan fogalmazzuk meg. A gyakorlatban akkor is alkalmazzák az *u*-próbát, amikor a minta nem normális eloszlású, de a mintaelemszám „nagy”. Az alkalmazás jogosságát a centrális határeloszlás-tétellel lehet indokolni. Ugyanis a próbastatisztika normális eloszlású lesz aszimptotikusan, mivel a CHT szerint a mintaátlag már közel normális eloszlású!

<table>
<tr><th colspan="2">Várható érték paraméter</th><th>Szórás paraméter</th></tr>
<tr><td colspan="2">egymintás u – próba</td><td rowspan="4">F – próba</td></tr>
<tr><td colspan="2">kétmintás u – próba</td></tr>
<tr><td colspan="2">egymintás t – próba</td></tr>
<tr><td rowspan="2">kétmintás t – próba</td><td>független mintás</td></tr>
<tr><td>összetartozó mintás</td><td rowspan="3">Bartlett-teszt</td></tr>
<tr><td colspan="2">Welch – próba</td></tr>
<tr><td colspan="2">egyszerű csoportosítás (one-way ANOVA)</td></tr>
</table>

Egymintás *u* – próba

Paraméter	Érték
Feltétel	A normális eloszlású mintának ismerjük a szórását.
H_0	A minta várható értéke m_0
Számolás	$u_{próba} = \frac{\overline{x}_n - m_0}{\sigma_0} \cdot \sqrt{n} \in N(0,1)$

	$P\left(\bar{x}_n - \frac{u_\varepsilon \cdot \sigma_0}{\sqrt{n}} < \vartheta < \bar{x}_n + \frac{u_\varepsilon \cdot \sigma_0}{\sqrt{n}}\right) = konfidencia\ szint$
Táblázat	Standard – normál $1 - konfidenciaszint = \varepsilon$ $\Phi(u_\varepsilon) = 1 - \frac{\varepsilon}{2}$ tartozó érték kétoldali esetben, H_0 –ra vonatkoztatott egyenlőség esetén $\Phi(u_\varepsilon) = 1 - \varepsilon$ egyoldali esetben, H_0 –ra vonatkoztatott egyenlőtlenség esetén
Döntés	$\lvert u_{próba} \rvert < u_\varepsilon \Rightarrow H_0$
Elsőfajú hiba	ε szignifikancia szint
Másodfajú hiba	$\frac{\overline{x}_n - m_0}{\sigma_0} \cdot \sqrt{n} \in N(0,1)$
Erő	Az u – próba tulajdonságai: a próba torzítatlan és konzisztens! Ráadásul egyenletesen legjobb próba is!
Megjegyzés	A gyakorlatban akkor is alkalmazzák az u – próbát, amikor a minta nem normális eloszlású, de a mintaelemszám „nagy”. Az alkalmazás jogosságát a centrális határeloszlás-tétellel lehet indokolni. Ugyanis a próbastatisztika normális eloszlású lesz aszimptotikusan, mivel a CHT szerint a mintaátlag már közel normális eloszlású!

Egymintás t – próba

Paraméter	Érték
Feltétel	A normális eloszlású mintának NEM ismerjük a szórását, így azt a minta realizációjából kell számolnunk
H_0	A minta várható értéke m_0
Számolás	$t_{próba} = \frac{\overline{x}_n - m_0}{s_n^*} \cdot \sqrt{n} \in t_{n-1}$ $P\left(\bar{x}_n - \frac{u_e \cdot s_n^*}{\sqrt{n}} < \vartheta < \bar{x}_n + \frac{u_e \cdot s_n^*}{\sqrt{n}}\right) = konfidencia\ szint$
Táblázat	Student – eloszlás $1 - konfidenciaszint = \varepsilon$ $\Phi(u_{krit}) = 1 - \frac{\varepsilon}{2}$ tartozó érték $n - 1$ szabadságfokkal
Döntés	$\lvert t_{próba} \rvert < t_{krit} \Rightarrow H_0$
Megjegyzés	Ha NEM ismerjük a szórást, de elég nagy a minta elemszám, akkor u – próbát alkalmazunk.

Kétmintás u – próba

Paraméter	Érték
Feltétel	Adottak az $X_1, X_2, X_3, \dots X_n$ és az $Y_1, Y_2, Y_3, \dots Y_m$ egymástól független statisztikai minták. A minták független normális eloszlásúak, a szórásaik ismertek.
H_0	$\mu_1 = \mu_2$
Számolás	$\overline{X}_n \in N\left(\mu_1, \frac{\sigma_1}{\sqrt{n}}\right), \quad \overline{Y}_m \in N\left(\mu_2, \frac{\sigma_2}{\sqrt{m}}\right)$ $\overline{X}_n - \overline{Y}_m \in N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right)$ $\overline{X}_n - \overline{Y}_m \in N\left(0, \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right)$ $u_{próba} = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \in N(0,1)$
Táblázat	Standard – normál $\Phi(u_\varepsilon) = 1 - \frac{\varepsilon}{2}$ tartozó érték kétoldali esetben, H_0 –ra vonatkoztatott egyenlőség esetén $\Phi(u_\varepsilon) = 1 - \varepsilon$ egyoldali esetben, H_0 –ra vonatkoztatott egyenlőtlenség esetén
Döntés	$\left\lvert \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \right\rvert < u_\varepsilon \Rightarrow H_0$

Kétmintás t – próba (független minták)

Paraméter	Érték
Feltétel	Adottak az $X_1, X_2, X_3, \dots X_n$ és az $Y_1, Y_2, Y_3, \dots Y_m$ egymástól független statisztikai minták. A minták független normális eloszlásúak, a szórásaik NEM ismertek. A minták szórásai egyenlőeknek tekintendők. Különben nem alkalmazható a próba. Ennek ellenőrzése *F*-próbával.
H_0	$\mu_1 = \mu_2$
Számolás	$t_{próba} = \frac{\overline{x}_n - \overline{y}_m}{\sqrt{(n-1)\cdot\left(s^*_{x,n}\right)^2 + (m-1)\cdot\left(s^*_{y,m}\right)^2}} \cdot \sqrt{\frac{n\cdot m\cdot(n+m-2)}{n+m}} \in t_{n+m-2}$ vagy átszámolni kétmintás u – próbára

	$\sigma^2 = \frac{(n-1)\cdot s_x^{*2} + (m+1)\cdot s_y^{x2}}{n+m-2}$
Táblázat	Student - eloszlás $\mathrm{P}(\lvert t_{n+m-2}\rvert < t_{krit}) = 1 - \frac{\varepsilon}{2}$ tartozó érték $n+m-2$ szabadságfokkal
Döntés	$\lvert t_{próba}\rvert < t_{krit} \Rightarrow H_0$

Kétmintás t – próba (összetartozó minták)

Paraméter	Érték
Feltétel	Adottak az $X_1, X_2, X_3, \dots X_n$ és az $Y_1, Y_2, Y_3, \dots Y_n$ egymástól NEM független statisztikai minták. A minták független normális eloszlásúak, a szórásaik NEM ismertek.
$\boldsymbol{H_0}$	$\mu_1 = \mu_2$ vagyis a minták várható értékei egyenlők
Számolás	$t_{próba} = \frac{\overline{x}_n - \overline{y}_n}{\sqrt{\left(s_{x,n}^*\right)^2 + \left(s_{y,n}^*\right)^2}} \cdot \sqrt{n} \in t_{2n-2}$ *vagy* Kivonjuk a két minta adatait egymásból és egymintás *t* – próbát alkalmazunk $T = \frac{\overline{z}_n}{s_z^*} \cdot \sqrt{n}$
Táblázat	Student - eloszlás $\mathrm{P}(\lvert t_{2n-2}\rvert < t_{krit}) = 1 - \frac{\varepsilon}{2}$ tartozó érték $2n-2$ szabadságfokkal
Döntés	$\lvert t_{próba}\rvert < t_{krit} \Rightarrow H_0$

F- próba

Paraméter	Érték
Feltétel	Adottak az $X_1, X_2, X_3, \dots X_n$ és az $Y_1, Y_2, Y_3, \dots Y_m$ egymástól független statisztikai minták. A minták független normális eloszlásúak, a szórásaik NEM ismertek.
$\boldsymbol{H_0}$	$\sigma_1 = \sigma_2$ vagyis a minták szórása egyenlők
Számolás	$\frac{\left(s_{x,n}^*\right)^2}{\left(s_{y,m}^*\right)^2} \in F_{n-1,m-1}$
Táblázat	*F* – táblázat

	$P(K_1 < F_{n-1,k-1}) = 1 - \frac{\varepsilon}{2}$ $n-1, m-1$ szabadságfokkal
Döntés	$\lvert t_{próba} \rvert < t_{krit} \Rightarrow H_0$

Nem paraméteres próbák

Ha az alapsokaság (a statisztikai minta) eloszlását nem tekintjük eleve ismertnek, akkor nemparaméteres próbákról beszélünk. Ilyenkor tehát az előzetes feltevéseink nagyon általánosak, de természetesek; pl. feltesszük, hogy a minta eloszlása folytonos, vagy feltesszük, hogy a szórás véges, stb. Mivel kevesebb feltételt követelünk meg kiinduláskor (a priori feltevések), a következtetéseink levonásához nagyobb elemszámú mintákra lesz szükségünk, mint a paraméteres próbák esetén. A próbastatisztikák eloszlását csak aszimptotikusan ismerjük.

Feladat	H_0	**Próbák**
Illeszkedésvizsgálat	Az elemzett változó eloszlása megegyezik a hipotetikussal	χ^2-próba
		egymintás Kolmogorov-Szmirnov
		P-P grafikon
Függetlenségvizsgálat	Az elemzett változók függetlenek	χ^2-próba
		nominális változókra
		ordinális változókra
Homogenitásvizsgálat	Az elemzett változók eloszlása azonos	χ^2-próba
		kétmintás Kolmogorov-Szmirnov
		Wilcoxon,
		McNemar
		Kruskal-Wallis
		Friedmann

χ^2-próba tiszta illeszkedésvizsgálat

Paraméter	**Érték**
Feltétel	Adott az $X_1, X_2, X_3, \ldots X_n$ statisztikai minta. $F_0(x)$ a minta hipotetikus eloszlásfüggvénye.
H_0	$T_n < K_\varepsilon$
Számolás	$T_n(X_1, X_2, \ldots, X_n) = \sum_{k=1}^{r} \frac{(\upsilon_k - n \cdot p_k)^2}{n \cdot p_k} \to \chi^2_{r-1} \quad (n \to \infty).$
Táblázat	χ^2– táblázat $\mathbf{P}(\chi^2_{r-1} < K_\varepsilon) = 1 - \varepsilon$

	$r-1$ szabadságfokkal
Döntés	$T_n < K_\varepsilon$

χ^2-próba becsléses illeszkedésvizsgálat

Paraméter	**Érték**
Feltétel	Adott az $X_1, X_2, X_3, \dots X_n$ statisztikai minta. $F_{\underline{\upsilon}}(x)$ a minta hipotetikus eloszlásfüggvénye. Az eloszlásfüggvény most k db paramétertől függ, aminek értékét nem ismerjük!
H_0	$T_n < K_\varepsilon$
Számolás	Csoportosítás $T_n(X_1, X_2, \dots, X_n) = \sum_{k=1}^{r} \frac{(\upsilon_k - n \cdot p_k)^2}{n \cdot p_k} \to \chi^2_{r-1} \quad (n \to \infty).$
Táblázat	χ^2 – táblázat $P(\chi^2_{r-1-k} < K_\varepsilon) = 1 - \varepsilon$ $r-1-k$ szabadságfokkal
Döntés	$T_n < K_\varepsilon$

χ^2-próba függetlenségvizsgálat

Paraméter	**Érték**
Feltétel	Adott az $(X_1, Y_1)^T, (X_2, Y_2)^T \dots (X_n, Y_n)^T$ n elemszámú kétdimenziós statisztikai minta. Ellenőrizni akarjuk, hogy a minta komponensei függetlenek-e egymástól, vagy pedig szignifikáns sztochasztikus összefüggés tapasztalható-e közöttük
H_0	$P(X_i < x, Y_i < y) = P(X_i < x)P(Y_i < y) \forall x, y$
Számolás	Csoportosítás Sorok és oszlopok összeadása Várt gyakoriságok számolása $k^*_{ij} = \frac{k_i \cdot k_j}{n}$ $T_n = n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(k_{ij} - k^*_{ij})^2}{k^*_{ij}}$
Táblázat	χ^2 – táblázat $P(\chi^2_{(r-1)(s-1)} < K_\varepsilon) = 1 - \varepsilon$ $(r-1)(s-1)$ szabadságfokkal
Döntés	$T_n < K_\varepsilon$

χ^2-próba homogenitásvizsgálat

Paraméter	Érték
Feltétel	A homogenitásvizsgálat annak a kérdésnek az eldöntésére szolgál, hogy két valószínűségi változó azonos eloszlású-e, azaz ugyanaz a függvény-e az eloszlásfüggvényük, vagy sem. Adottak az $X_1, X_2, X_3, \dots X_n$ és az $Y_1, Y_2, Y_3, \dots Y_m$ statisztikai minták, amelyek egymástól függetlenek
H_0	$P(X < x) \equiv P(Y < y)$
Számolás	Csoportosítás Táblázat esetén sorok (v_i), oszlopok (μ_i) összeadása $T = n_1 \cdot n_2 \sum_{i=1}^{r} \frac{\left(\frac{v_i}{n_i} - \frac{\mu_i}{n_2}\right)^2}{v_i + \mu_i}$
Táblázat	χ^2 – táblázat $\mathbf{P}(\chi^2_{r-1} < K_\varepsilon) = 1 - \varepsilon$ $f = r - 1$ szabadságfokkal
Döntés	$T_n < K_\varepsilon$

Kolmogorov-Szmirnov – próba (homogenitásvizsgálat)

Paraméter	Érték
Feltétel	A homogenitásvizsgálat annak a kérdésnek az eldöntésére szolgál, hogy két valószínűségi változó azonos eloszlású-e, azaz ugyanaz a függvény-e az eloszlásfüggvényük, vagy sem. Adottak az $X_1, X_2, X_3, \dots X_n$ és az $Y_1, Y_2, Y_3, \dots Y_m$ statisztikai minták, amelyek egymástól függetlenek
H_0	A minta eloszlásfüggvénye $F(x)$
Számolás	$t_{próba} = \sqrt{n} \cdot sup\left\|F_{emp}(x) - F(x)\right\|$
Táblázat	Kolmogorov – Szmirnov – táblázat $K(\chi_\varepsilon) = 1 - \varepsilon$
Döntés	$\left\|t_{próba}\right\| < t_{krit} \Rightarrow H_0$

$n \cdot p_i$	várható érték
n	kísérletek száma
p_i	i – edik esemény bekövetkezési valószínűsége
r, s	csoportok száma
k	paraméterek száma

Maximum likelihood becslés

A módszer alapgondolatai a következők:

- A mintánk eloszlásfüggvénye a υ paramétertől függ.
- Ha egy kísérletnél több esemény is bekövetkezhet, legtöbbször a legnagyobb valószínűségű eseményt fogjuk megfigyelni.
- A sokaságra vett mintavételezés során kaptunk egy realizációt. Feltételezzük, hogy azért éppen ezt a realizációt kaptuk, és nem mást, mert az összes realizációk közül ennek volt a legnagyobb a bekövetkezési valószínűsége.
- Vegyük tehát, az összes lehetséges υ paraméter közül azt, amelynél éppen kapott realizáció bekövetkezése a maximális.

Legyen adott egy P, valószínűségi mértékek egy tere és az $X_1, X_2, X_3, \dots X_n$ diszkrét eloszlású statisztikai minta $E \in P$-re.

$$L(x,\vartheta) = P_\vartheta(X_1 = x_1, X_2 = x_2, X_3 = x_3, \dots X_n = x_n) = \prod_{i=1}^{n} P_\vartheta(X_i = x_i)$$

a minta együttes eloszlását. A ϑ paraméter maximum – likelihood becslésén azt a $\tau_n(X_1, X_2, X_3, \dots X_n)$ statisztikát értjük, melyre

$$L\big(x, \tau_n(x)\big) = \max_{\vartheta \in \mathbb{R}^k} L(x,\vartheta)$$

teljesül.

Számolás

1.	Vegyük a minta eloszlás, pl.: Poisson – eloszlás	$p_{\vartheta,i} = \frac{\vartheta^i}{i!} \cdot e^{-\vartheta}$, $i = 0, 1, 2 \dots$
2.	Számoljuk a likelihood függvényt, a minta együttes eloszlásából.	$L(x,\vartheta) = \prod_{i=1}^{n} \left(\frac{\vartheta^{x_i}}{x_i!} \cdot e^{-\vartheta} \right) = \frac{\vartheta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} \cdot e^{-n\cdot\vartheta}$
3.	Számoljuk a log – likelihood függvényt, mely nem változtatja meg a maximumhelyeket.	$l(x,\vartheta) = ln\vartheta \cdot \sum_{i=1}^{n} x_i - n \cdot \vartheta - ln\left(\prod_{i=1}^{n} x_i! \right)$
4.	Stacionárius helyek megkeresése.	$\frac{\partial l(x,\vartheta)}{\partial \vartheta} = \frac{1}{\vartheta} \sum_{i=1}^{n} x_i - n = 0 \rightarrow \vartheta = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}_n$
5.	Visszaellenőrizni, hogy ezek milyen szélsőértékhelyek.	$\frac{\partial^2 l(x,\vartheta)}{\partial \vartheta^2} = -\frac{1}{\vartheta^2} \sum_{i=1}^{n} x_i < 0$

A kapott stacionárius hely maximum, tehát a Poisson-eloszlás esetén is a paraméternek maximum – likelihood becslése az átlagstatisztika.

Általános feltételek mellett megmutatható, hogy a maximum – likelihood becslés konzisztens, aszimptotikusan normális eloszlású, és ha van elégséges statisztika, akkor a maximum likelihood statisztika éppen azt adja meg!

Momentumok módszere

Eloszlás	1. momentum
Bernoulli-eloszlás	$EX_1 = N \cdot \vartheta \approx \overline{x}_n \Rightarrow \vartheta = \frac{\overline{x}_n}{N}$
Poisson – eloszlás	$EX_1 = \vartheta \approx \overline{x}_n \Rightarrow \vartheta \approx \overline{x}_n$
Egyenletes - eloszlás	$EX_1 = \frac{1}{\vartheta} \approx \overline{x}_n \Rightarrow \vartheta \approx \frac{1}{\overline{x}_n}$
Normál - eloszlás	$EX_1 = \vartheta \approx \overline{x}_n \Rightarrow \vartheta \approx \overline{x}_n$

Hibák lehetnek benne!

Készítette: Horváth Gábor

Ellenőrzésben segítségemre volt: Fischer Hanna