A Valószínűségszámítás II. előadássorozat tizedik előadása.

2002. november 19.

A χ^2 próba és néhány egyéb érdekes statisztikai alkalmazás.

Idézzük fel azt a kérdést, melyet felvetettünk mint olyan problémát, melynek megoldásához érdemes bebizonyítani a több-dimenziós centrális határeloszlástételt.

> Egy dobókockáról el akarjuk dönteni, hogy szabályos-e. Ennek érdekében a dobókockát sokszor egymástól függetlenül feldobjuk, és feljegyezzük a kisérletek eredményét. Milyen eredménysorozat esetén tekinthetjük a dobókockát szabályosnak?
>
> Általánosabban:
>
> Legyen adva k urna, és ezekbe egymástól függetlenül bedobunk n golyót egymástól függetlenül, amelyek ugyanolyan valószínűséggel esnek az egyes urnákba. Próbáljuk meg ellenőrízni azt, hogy igaz-e, hogy a golyók az egyes urnákba $p_1, \dots, p_k$ valószínűséggel esnek. (Feltesszük, hogy az elvégzett kisérletek n száma nagyon nagy.)

Emlékeztettünk arra, hogy a feladat megoldásában azt érdemes vizsgálni, hogy a j-ik urnába eső golyók száma minusz azok várható értéke mennyire kicsi, és a több-dimenziós centrális határeloszlástétel segítségével azt kívánuk vizsgálni, hogy mennyire valószínű egy adott nagyságú eltérés. Annak érdekében, hogy a több-dimenziós centrális határeloszlástételt alkalmazhassuk érdemes bevezetni a $Z_j = (Z_j^{(1)}, \dots, Z_j^{(k)})$, $j = 1, 2, \dots, n$, véletlen vektorokat, melyeket a következő módon definiálunk: A Z_j vektor $Z_j^{(l)}$ koordinátája 1, és az összes többi koordinátája nulla, ha a j-ik dobásban az l-ik urnába esik a golyó. Ekkor a $Z_1, \dots, Z_n$ véletlen vektorok függetlenek, és ezek $S_n = (S_n^{(1)}, \dots, S_n^{(k)})$, $S_n = Z_1 + \dots + Z_n$ összege megadja, hogy hány pont esik az egyes urnákba.

Ezért a több-dimenziós centrális határeloszlástétel alapján ki tudjuk számolni a $\lim_{n\to\infty} P(S_n^{(1)} < ES_n^{(1)} + \sqrt{n}x_1, \dots, S_n^{(k)} < ES_n^{(k)} + \sqrt{n}x_k)$ valószínűségek határértékét. Viszont természetes módon felmerülhet a kérdés: Ki tudjuk-e számítani hasonló módon annak aszimptotikus valószínűségét, hogy az $S_n = (S_n^{(1)}, \dots, S_n^{(k)})$ vektor koordinátáinak megfelelő normalizáltja valamely ,,általános szép halmazba" essen. Be fogjuk látni, hogy erre a kérdésre pozitív választ lehet adni. De e válasz megadása előtt idézzük fel az alábbi eredményt, mely ezen eredmény bizonyításának alapjául szolgál.

Tétel az eloszlásban való konvergencia jellemzéséről folytonos függvények segítségével. *$F_n(u_1, \dots, u_k)$, $n = 1, 2, \dots$ k-dimenziós eloszlásfüggvények sorozata akkor és csak akkor konvergál eloszlásban egy $F(u_1, \dots, u_k)$ k-dimenziós eloszlásfüggvényhez, ha minden a k-dimenziós téren értelmezett folytonos, korlátos $g(u_1, \dots, u_k)$ függvényre teljesül a*

$$\lim_{n\to\infty} \int g(u_1, \dots, u_k)\, F_n(\, du_1, \dots, \, du_k) = \int g(u_1, \dots, u_k)\, F(\, du_1, \dots, \, du_k)$$

azonosság.

Most megfogalmazom azt az eredményt, mely pozitív választ ad az előző kérdésre.

Tétel. *Legyen $F_n(x_1,\dots,x_k)$, $n=1,2,\dots$, k-dimenziós eloszlásfüggvények sorozata, amelyik eloszlásban konvergál egy $F_n(x_1,\dots,x_k)$ k-dimenziós eloszlásfüggvényhez, $n\to\infty$ esetben. Jelölje μ_{F_n} az $F_n(x_1,\dots,x_k)$, $n=1,2,\dots$, és μ_F az $F(x_1,\dots,x_k)$, k-dimenziós eloszlásfüggvények által meghatározott Lebesgue–Stieltjes mértéket. Ekkor tetszőleges olyan (Borel-mérhető) $A\subset R^k$ halmazra a k-dimenziós térben, melynek ∂A határa 0 μ_F mértékű, azaz $\mu_F(\partial A)=0$, teljesül a $\lim_{n\to\infty}\mu_{F_n}(A)=\mu_F(A)$ reláció. (Egy A halmaz határán azokat a pontokat értjük, melyek torlódási pontjai mind az A halmaznak, mind annak komplementerének az $R^k\setminus A$ halmaznak.) Speciálisan, ha F egy k-dimenziós normális eloszlás nem elfajuló kovariancia mátrix-szal, akkor ez az állítás igaz minden olyan A halmazra, melynek a határa nulla Lebesque mértékű.*

Bizonyítás. A bizonyítás lényege az, hogy az A halmaz indikátor függvényét közrefogjuk olyan folytonos függvényekkel, mely függvényeknek a μ_{F_n} mérték szerinti integrálja jól közelíti az A halmaz $\mu_{F_n}(A)$ mértékét, és melyekre alkalmazhatjuk az eloszlásban való konvergenciának a folytonos függvények integráljával fent megadott jellemzését.

Rögzítsünk egy $\varepsilon>0$ számot, definiáljuk az A halmaz „ε-belsejét", mint a következő A^ε halmazt: $A^\varepsilon=\{x\colon \rho(x,R^k\setminus A)\ge\varepsilon\}$, ahol $\rho(x,y)$ két pont euklidészi távolságát jelöli az R^k euklidészi térben, és egy $x\in R^k$ pontra, és $B\subset R^k$ halmaz távolságát a szokásos módon a $\rho(x,B)=\inf_{y\in B}\rho(x,y)$ formulával definiáljuk. Definiáljuk az A halmaz ε-környezetét mint a következő B^ε halmazt: $B^\varepsilon=\bigcup_{x\in A}\{y\colon \rho(x,y)<\varepsilon\}$. Vegyük észre, hogy $A^\varepsilon\subset A\subset B^\varepsilon$, továbbá $\lim_{N\to\infty}\mu_F(A^{1/N})=\mu_F(A)$ és $\lim_{N\to\infty}\mu_F(B^{1/N})=\mu_F(A)$. Valóban, az $A^{1/N}$ halmazok egymásba skatulyázottak, ezért $\lim_{N\to\infty}\mu_F(\lim_{N\to\infty}A^{1/N})=\mu_F\left(\bigcup_{N=1}^{\infty}A^{1/N}\right)=\mu_F(\operatorname{Int}A)=\mu_F(A)$, ahol $\operatorname{Int}A=A\setminus\partial A$ az A halmaz belseje. Itt kihasználtuk, hogy $\mu_F(\partial A)=0$. Hasonlóan, $\lim_{N\to\infty}\mu_F(\lim_{N\to\infty}B^{1/N})=\mu_F\left(\bigcap_{N=1}^{\infty}B^{1/N}\right)=\mu_F(\bar A)=\mu_F(A)$, ahol $\bar A=A\cup\partial A$ az A halmaz lezártja.

Adva egy $B\in R^k$ halmaz, jelölje $I_B(x)$ a B halmaz indikátor függvényét, azaz legyen $I_B(x)=1$, ha $x\in B$, $I_B(x)=0$, ha $x\notin B$. Minden $\varepsilon>0$ számra tudunk olyan $f_\varepsilon(x)$ és $g_\varepsilon(x)$ folytonos függvényt konstruálni, melyre $I_{A^\varepsilon}(x)\le f_\varepsilon(x)\le I_A(x)\le g_\varepsilon(x)\le I_{B^\varepsilon}(x)$. Valóban, legyen $f_\varepsilon(x)=\frac{1}{\varepsilon}\min(\varepsilon,\rho(x,R^k\setminus A))$. Ekkor $f_\varepsilon(x)$ folytonos függvény, $f_\varepsilon(x)=0$, ha $x\notin A$, $f_\varepsilon(x)=1$, ha $x\in A^\varepsilon$, és $0\le f(x)\le 1$, ha $x\in A\setminus A^\varepsilon$. Ez azt jelenti, hogy $I_{A^\varepsilon}(x)\le f_\varepsilon(x)\le I_A(x)$ minden $x\in R^k$ pontban. Hasonlóan, a $g_\varepsilon(x)=\frac{1}{\varepsilon}\min(\varepsilon,\rho(x,R^k\setminus B^k)$ függvény is folytonos, és $I_A(x)\le f_\varepsilon(x)\le I_{B^\varepsilon}(x)$ minden $x\in R^k$ pontban. Ezért az előbb felidézet tétel alapján felírhatjuk, hogy

$$\begin{aligned}\mu_F(A^{1/N})&\le\int f_{1/N}(x)\mu_F(\,dx)=\lim_{n\to\infty}\int f_{1/N}(x)\mu_{F_n}(\,dx)\le\liminf_{n\to\infty}\mu_{F_n}(A)\\&\le\limsup_{n\to\infty}\mu_{F_n}(A)\le\lim_{n\to\infty}\int g_{1/N}(x)\mu_{F_n}(\,dx)=\int g_{1/N}(x)\mu_F(\,dx)\end{aligned}$$

$$\leq \mu_F(B^{1/N}).$$

Innen $N \to \infty$ határátmenettel kapjuk, felhasználva a $\lim_{N\to\infty} \mu_F(A^{1/N}) = \mu_F(A)$ és $\lim_{N\to\infty} \mu_F(B^{1/N}) = \mu_F(A)$ relációkat, hogy

$$\mu_F(A) \leq \liminf_{n\to\infty} \mu_{F_n}(A) \leq \limsup_{n\to\infty} \mu_{F_n}(A) \leq \mu_F(A).$$

Ebből az egyenlőtlenségből következik a Tétel fő állítása. A Tétel második állítása azonnal következik az elsőből, ha észrevesszük, hogy egy nem elfajuló kovarianci mátrixszal rendelkező normális eloszlásnak van sűrűségfüggvénye, ezért, ha egy A halmaz határának Lebesque mértéke nulla, akkor a határ mértéke nulla egy nem elfajululó kovariancia mátrixú normális eloszlás szerint is.

A fenti tétel elvileg jól alkalmazható, de biznoyos esetekben érdemes azt a halmazt, melynek a mértékét keressük a normális határeloszlásmérték szerint jobban megválasztani, olyan halmazoknak a valószínűségét vizsgálni, melyek ,,jobban illeszkednek a feladathoz", és megvizsgálni, hogy ebben az esetben nem tudjuk-e a határmértéket jobban, egyszerűbben leírni. Ilyen például az előadás elején felidézett probléma, melynek megoldásában általában az úgynevezett χ^2 (ejtsd khi négyzet) próbát szokták alkalmazni. Bár ennek tárgyalását általában a matematikai statisztika tananyag részének tekintik, a bizonyítás lényege a több-dimenziós centrális határeloszlástételen, illetve bizonyos lineáris algebrai ismereteken alapul. Ezért természetes ezt itt tárgyalni. Először ismertettem a χ^2 eloszlás fogalmát és megfogalmazom az eredményt.

A k szabadságfokú χ^2 eloszlás definiciója. *Legyen $\xi_1, \ldots, \xi_k$ k darab független standard normális eloszlású valószínűségi változó. Ekkor a $\sum_{j=1}^{k} \xi_j^2$ valószínűségi változó eloszlását nevezzük k szabadságfokú $\chi^2(k)$ eloszlásnak.*

Megjegyzés: Láttuk a gyakorlaton vett feladatok egyikében, hogy a 2 szabadságfokú $\chi^2(2)$ eloszlás a $\lambda = \frac{1}{2}$ paraméterű exponenciális eloszlás, azaz az az eloszlás, melynek sűrűségfüggvénye az $f(x) = \frac{1}{2}e^{-x/2}$, ha $x \geq 0$, és $f(x) = 0$, ha $x < 0$.

Most megfogalmazom a tárgyalandó eredményt.

A χ^2 próbáról szóló tétel. *Legyen adva k darab urna, melyekbe bedobunk egymástól függetlenül golyókat úgy, hogy mindegyik golyó p_j valószínűséggel esik a j-ik urnába, $1 \leq j \leq k$, $\sum_{j=1}^{k} p_j = 1$. Jelölje $\nu_n(j)$ a j-ik urnába eső golyók számát az n-ik dobás után. Ekkor a $\sum_{j=1}^{k} \frac{(\nu_n(j) - np_j)^2}{np_j}$ valószínűségi változók eloszlásban konvergálnak a $k-1$ szabadságfokú $\chi^2(k-1)$ eloszláshoz, ha $n \to \infty$. (Az urnák k száma rögzített.)*

1. megjegyzés: A fenti tételben megjelenő határeloszlás csak az urnák k számától függ, de nem függ a p_j, $1 \leq j \leq k$, valószínűségektől. Ez jelzi azt, hogy természetes statisztikát vettünk, olyat amelyben a különböző urnákban levő golyók számának az eltérése annak várható értékétől egyforma fontos szerepet játszik.

2. megjegyzés: Az, hogy a határeloszlás a $k-1$ szabadságfogú $\chi^2(k-1)$ eloszlás azzal függ össze, hogy bár k véletlen szám súlyozott négyzetösszegét tekintettük, (az egyes urnákba eső golyók számának eltérését tekintettük azok várható értékétől), de ezek között van egy determinisztikus összefüggés. Nevezetesen az, hogy az összes urnába eső golyók száma minusz azok várható értéke nullával egyenlő. Ezt informálisan úgy szokták interpretálni, hogy mivel a k változó között volt egy összefüggés, ezért a vektor koordinátáinak szabadságfoka eggyel csökkent, és csak $k-1$ szabadsági fokkal rendelkező véletlen vektorok koordinátáinak a négyzetösszegét tekintettük, illetve azok határeloszlását. Ilyen esetben a határeloszlást olyan véletlen összeg adja meg, melyben mindegyik szabadsági foknak egy összeadandó felel meg, amelyik független a többi összeadandótól, és egy standard normális eloszlású valószínűségi változó négyzete.

Először megfogalmazok és bebizonyítok egy állítást, mely lehetővé teszi, hogy a χ^2 próbáról szóló tételt redukáljuk egy olyan állításra, melyben egy alkalmaz kovarianciájú Gauss eloszlású véletelen vektor koordinátáinak a négyzetösszegét határozzuk meg.

Tétel. *Legyen $(S_{1,n}, \ldots, S_{k,n})$, $n = 1, 2 \ldots$, k-dimenziós valószínűségi vektorok sorozata, amelyik eloszlásban konvergál egy $(S_1, \ldots, S_k)$ k-dimenziós véletlen vektorhoz $n \to \infty$ esetén, és legyen $f(x_1, \ldots, x_k)$ egy k-változós folytonos függvény. Ekkor a $T_n = f(S_{1,n}, \ldots, S_{k,n})$, $n = 1, 2, \ldots$, valószínűségi változók eloszlásban konvergálnak a $T = f(S_1, \ldots, S_k)$ valószínűségi változóhoz $n \to \infty$ esetén.*

Bizonyítás: Használjuk a tételt az eloszlásban való konvergencia jellemzéséről folytonos függvények segítségével. Eszerint az $S_{1,n}, \ldots, S_{k,n}$ véletlen vektorok sorozatának eloszlásban való konvergenciát úgy is megfogalmazhatjuk, hogy tetszőleges folytonos és korlátos $h(x_1, \ldots, x_k)$ függvényre teljesül a $\lim_{n\to\infty} Eh(S_{1,n}, \ldots, S_{k,n}) = Eh(S_1, \ldots, S_k)$ reláció. Legyen $g(\cdot)$ tetszőleges folytonos és korlátos függvény a számegyenesen. Ekkor $g(f(x_1, \ldots, x_k))$ is folytonos és korlátos függvény, ezért teljesül a

$$\lim_{n\to\infty} Eg(T_n) = \lim_{n\to\infty} Eg\left(f\left(S_{1,n}, \ldots, S_{k,n}\right)\right) = Eg\left(f\left(S_1, \ldots, S_k\right)\right) = Eg(T)$$

reláció. Innen következik a Tétel állítása.

Megjegyzés: Az előbb kimondott tétel valójában sokkal általánosabb körülmények között érvényes. Lehet definiálni általános (szeparábilis metrikus tér) értékű valószínűségi változókat, azok eloszlását és eloszlásaik konvergenciáját. A definiciók eléggé természetesek, és a fenti eredmény általános terekben is érvényes, sőt a bizonyítást sem kell megváltoztatni. Ennek az állításnak a részleteire nem térek ki. Azt érdemes megjegyezni, hogy ezek az állítások nem pusztán formális általánosítások, hanem nagyon érdekes és hasznos és tartalmas következményei vannak a ,,mindennapi" valószínűségi változók vizsgálatában is.

Most megfogalmazom, hogyan lehet redukálni a a χ^2 próbáról szóló tételt.

A χ^2 próbáról szóló tétel redukciója. *Legyen $\eta = (\eta_1, \dots, \eta_k)$ k-dimenziós normális eloszlású véletlen vektor, melynek várható értéke és kovariancia mátrixa teljesíti az $E\eta_j = 0$, $1 \le j \le k$, $\mathrm{Cov}\,(\eta_j, \eta_l) = -\sqrt{p_j p_l}$, $1 \le j, l \le k$, $j \ne l$, és $\mathrm{Var}\,\eta_j = (1 - p_j)$, $1 \le j \le k$ relációkat. Akkor $\sum_{j=1}^{k} \eta_j^2$ eloszlása a $k-1$ szabadságfokú $\chi^2(k-1)$ eloszlás. (A most megfogalmazott eredményben speciálisan azt is állítjuk, hogy létezik az adott kovarianciával rendelkező véletlen vektor tetszőleges p_j, $p_j \ge 0$, $1 \le j \le k$, $\sum_{j=1}^{k} p_j = 1$ számokra.)*

A χ^2 próbáról szóló tétel visszavezetése annak redukciójára. Ahogy az előadás elején tetük, vezessük be a $Z_m = (Z_m^{(1)}, \dots, Z_m^{(k)})$, $m = 1, 2, \dots, n$, véletlen vektorokat, melyeket a következő módon definiálunk: A Z_m vektor $Z_m^{(j)}$ koordinátája 1, és az összes többi koordinátája nulla ha az m-ik dobásban az j-ik urnába esik a golyó. Vezessük be ennek a következő transzformáltjait: $X_m = (X_m^{(1)}, \dots, X_m^{(k)})$, $m = 1, 2, \dots, n$, $X_m^{(j)} = \dfrac{Z_m^{(j)} - p_j}{\sqrt{p_j}}$, $1 \le m \le n$, $1 \le j \le k$. Vegyük észre egyrészt, hogy a vizsgált $\nu_n(j)$ mennyiségek (a j-ik urnába eső golyók száma az n-ik dobás után és a fenti X_m véletlen vektorok az alábbi azonosságot teljesítik: $\dfrac{1}{\sqrt{n}} \sum_{l=1} X_m = \left(\dfrac{\nu_n(1) - np_1}{\sqrt{np_1}}, \dots, \dfrac{\nu_n(k) - np_k}{\sqrt{np_k}} \right)$. Továbbá az X_m vektorok függetlenek és egyforma eloszlásúak, várható érték vektoruk nulla, és kovariancia mátrixuk teljesíti a $\mathrm{Cov}\,(X_m^{(j)}, X_m^{(l)}) = -\sqrt{p_j p_l}$, $1 \le j, l \le k$, $j \ne l$, és $\mathrm{Var}\, X_m^{(j)} = (1 - p_j)$, $1 \le j, l \le k$, $1 \le m \le n$ relácókat. Ugyanis egyszerű számolás adja, hogy $EZ_m = (p_1, \dots, p_k)$, $\mathrm{Cov}\,(Z_m^{(j)}, Z_m^{(l)}) = -p_j p_l$, $1 \le j, l \le k$, $j \ne l$, és $\mathrm{Var}\, Z_m^{(j)} = p_j(1 - p_j)$, $1 \le j, l \le k$, $1 \le m \le n$. Innen látszik, hogy $EX_m = 0$, és $\mathrm{Cov}\,(X_m^{(j)}, X_m^{(l)}) = \dfrac{\mathrm{Cov}\,(Z_m^{(j)}, Z_m^{(l)})}{\sqrt{p_j p_l}} = -\sqrt{p_j p_l}$, $\mathrm{Var}\, X_m^{(j)} = \dfrac{\mathrm{Var}\, Z_m^{(j)}}{p_j} = 1 - p_j$. Innen következik speciálisan az is, hogy a χ^2 próbáról szóló tétel redukciójában szereplő kovarianca mátrixú (nulla várható értékú) (normális eloszlású) $(\eta_1, \dots, \eta_k)$ véletlen vektor valóban létezik. Továbbá, a több-dimenziós centrális határeloszlástétel szerint a $\left(\dfrac{\nu_n(1) - np_1}{\sqrt{np_1}}, \dots, \dfrac{\nu_n(k) - np_k}{\sqrt{np_k}} \right)$ eloszlású véletlen vektorok eloszlásban konvergálnak egy ilyen$(\eta_1, \dots, \eta_k)$ normális vektorhoz. Ezért alkalmazva az előző tételt az $f(x_1, \dots, x_k) = \sum_{j=1}^{k} x_j^2$ folytonos függvénnyel kapjuk, hogy a $\sum_{j=1}^{k} \dfrac{(\nu_n(j) - np_j)^2}{np_j}$ valószínűségi változóknak van határeloszlásuk, és az megegyezik a $\sum_{j=1}^{k} \eta_j^2$ valószínűségi változó eloszlásával. Ez azt jelenti, hogy a χ^2 próbáról szóló tétel bizonyításához elég annak redukált változatát belátni.

A fenti eredmény azt jelenti, hogy egy normális eloszlású véletlen vektor koordi-

nátáiból képezett kvadratikus forma eloszlását kell kiszámolnunk. Emlékezzünk arra, hogy egy normális vektor lineáris transzformáltja is normális eloszlású, és tetszőleges normális eloszlású vektor előáll, mint standard normális eloszlású véletlen vektor lineáris transzformáltja. Ez azt sugallja, hogy próbáljuk meg a vizsgált (vagy egy vele azonos eloszlású) kifejezést előállítani, mint alkalmas független standard normális eloszláaú valószínűségi változók egyszerű kvadratikus formáját. A χ^2 próbáról szóló tétel redukciójában eredetileg megadott kifejezést azért kívánjuk átalakítani, mert igaz ugyan, hogy normális valószínűségi változók nagyon egyszerű kvadratikus formájáról van szó, viszont a kvadratikus formában szereplő változók nem függetlenek. Viszont az alábbi lemma nagyon hasznos lesz a számunkra.

Lemma. *Legyen $\eta = (\eta_1, \dots, \eta_k)$ k-dimenziós normális eloszlású valószínűségi változó nulla várható értékkel és D kovariancia mátrix-szal. Legyenek a D mátrix sajátértékei a $\lambda_1, \dots, \lambda_k$ számok (multiplicitással). Ekkor a $\sum_{j=1}^{k} \eta_j^2$ valószínűségi változó eloszlása megegyezik egy $\sum_{j=1}^{k} \lambda_j \xi_j^2$ valószínűségi változó eloszlásával, ahol $\xi_1, \dots, \xi_k$ független standard normális eloszlású valószínűségi változók.*

Bizonyítás: A D mátrix felírható $D = U\Lambda U^*$ alakban, ahol U unitér, Λ pedig olyan diagonális mátrix, melynek átlójában a D mátrix λ_j sajátértékei vannak (multiplicitással). (Az U mátrix is felírható explicit módon a D nátrix sajátvektorainak segítségével, de erre a tényre most nincs szükségünk.) Az $\eta = (\eta_1, \dots, \eta_k)$ véletlen vektor eloszlása megegyezik egy $\bar{\eta} = (\bar{\eta}_1, \dots, \bar{\eta}_k) = \xi\Lambda^{1/2}U^* = (\xi_1, \dots, \xi_k)\Lambda^{1/2}U^*$ véletlen vektor eloszlásával, ahol $\xi = (\xi_1, \dots, \xi_k)$ standard normális eloszlású véletlen vektor. Valóban $\bar{\eta}$ normális eloszlású véletlen vektor, melynek várható értéke nulla és kovariancia mátrixa a $(\Lambda^{1/2}U^*)^*\Lambda^{1/2}U^* = U\Lambda^{1/2}\Lambda^{1/2}U^* = U\Lambda U^* = D$ mátrix. Ezért az η és $\bar{\eta}$ véletlen vektorok eloszlása megegyezik. Innek az is következik, hogy a $\sum_{j=1}^{k} \eta_j^2$ valószínűségi változó eloszlása megegyezik a $\sum_{j=1}^{k} \bar{\eta}_j^2$ valószínűségi változó eloszlásával. Vegyük észre, hogy az $\tilde{\eta} = (\tilde{\eta}_1, \dots, \tilde{\eta}_k) = \bar{\eta}U = (\bar{\eta}_1, \dots, \bar{\eta}_k)U$ vektorra teljesül a $\sum_{j=1}^{k} \bar{\eta}_j^2 = \sum_{j=1}^{k} \tilde{\eta}_j^2$ azonosság, mert U unitér, tehát távolságtartó transzformáció. Viszont $\tilde{\eta} = \bar{\eta}U = \xi\Lambda^{1/2}U^*U = \xi\Lambda^{1/2}$. Ez azt jelenti, hogy a $\sum_{j=1}^{k} \eta_j^2$ valószínűségi változó eloszlása megegyezik a $\sum_{j=1}^{k} (\lambda_j^{1/2}\xi_j)^2 = \sum_{j=1}^{k} \lambda_j \xi_j^2$ valószínűségi változó eloszlásával, és ez a Lemma állítása.

Most befejezzük a χ^2 próbáról szóló tétel bizonyítását.

A χ^2 próbáról szóló tétel redukciójának a bizonyítása. Tekintsük a $\sum_{j=1}^{k} \eta_j^2$ valószínűségi

változót, ahol $(\eta_1, \dots, \eta_k)$ k-dimenziós normális eloszlású valószínűségi változó, melynek várható értéke nulla, és $D = d_{j,l}$, $d_{j,l} = \mathrm{Cov}\,(\eta_j, \eta_l)$, $1 \le j, l \le k$, kovariancia mátrixát a $\mathrm{Var}\,\eta_j = (1 - p_j)$, $1 \le j, l \le k$, és $\mathrm{Cov}\,(\eta_j, \eta_l) = -\sqrt{p_j p_l}$, ha $1 \le j, l \le k$, és $j \ne l$ képletek határozzák meg. Az előző lemma alapján azt kell megmutatni, hogy a fenti D kovariancia mátrixnak az 1 $k-1$ multiplicitású sajátértéke (azaz $k-1$ ortonormált 1 sajátértékkel rendelkező sajátvektora van) és ezenkívül még a nulla a sajátértéke 1-szeres multiplicitással.

Írjuk fel a D mátrixot $D = I - B$ alakban, ahol I az identitás mátrix, $B = (b_{i,j})$, $b_{i,j} = \sqrt{p_i}\sqrt{p_j}$, $1 \le i, j \le k$, és vegyük észre, hogy amennyiben egy B mátrixnak a sajátvektorai $e_1, \dots, e_k$ vektorok, $\lambda_1, \dots, \lambda_k$ sajátértékkel, akkor tetszőleges c számra az $I + cB$ mátrix sajátvektorai ugyanazok az $e_1, \dots, e_k$ vektorok $1 + c\lambda_1, \dots, 1 + c\lambda_k$ sajátértékkel. Ezt az eredményt alkalmazva $c = -1$ választással elegendő megtalálni a B mátrix sajátértékeit. Viszont egy $B = (b_i b_j)$, $1 \le i, j \le k$ alakú mátrix egyik sajátvektora a $b = (b_1, \dots, b_k)$ vektor $\sum_{j=1}^{k} b_j^2$ sajátértékkel, és a b vektort ortogonálisan kiegészítő altér a B mátrix $k-1$-dimenziós saját altere nulla sajátértékkel. Ez azt jelenti, hogy a nulla a B mátrix $k-1$ multiplicitású sajátértéke, mert B $k-1$-dimenziós nulla sajátértékű sajátalterének egy ortonormált bázisa $k-1$ ortonormált sajátvektort biztosít nulla sajátértékkel.

Az előbb megfogalmazott állítások egyszerűen ellenőrizhetőek. Valóban egyszerű számolás mutatja, hogy a $b = (b_1, \dots, b_k)$ vektorra $bB = \left(\sum_{j=1}^{k} b_j^2\right) b$, és ha valamely $c = (c_1, \dots, c_k)$ vektorra $\sum_{j=1}^{k} c_j b_j = 0$, akkor $cB = 0$, mert e vektor l-ik koordinátája $b_l \sum_{j=1}^{k} c_j b_j = 0$, és ez jelenti a b vektor ortogonális kiegészítő alterére megfogalmazott állítást.

Jelen esetben a tekintett B mátrix a fent vizsgált alakú a $b_j = \sqrt{p_j}$ számokkal. Ezért $\sum_{j=1}^{k} b_j^2 = \sum_{j=1}^{k} p_j = 1$, a B vektornak a nulla $k-1$-szeres az 1 pedig egyszeres multiplicitású sajátértéke. Ez azt jelenti, hogy a $D = I - B$ mátrixnak a nulla egyszeres az 1 pedig $k-1$-szeres multiplicitású gyöke, mint állítottuk.

Végül bebizonyítunk még egy a több-dimenziós normális eloszlás most megismert tulajdonságai alapján egyszerűen bizonyítható, és a matematikai statisztikában fontos szerepet játszó eredményt.

Tétel. *Legyenek $X_1, \dots, X_k$ független standard eloszlású valószínűségi változók, és legyen $\bar{X} = \frac{1}{k}\sum_{j=1}^{k} X_j$, $\bar{X}_j = X_j - \bar{X}$. Definiáljuk az $U = \sqrt{k}\bar{X}$ és $V = \sum_{j=1}^{k} \bar{X}_j^2$ valószínűségi változókat. Ekkor U és V független valószínűségi változók, továbbá U standard normális és V $\chi^2(k-1)$ eloszlású.*

Megjegyzés: A matematikai statisztikában vizsgálják a következő problémát. Tekint-

sük $(\xi_1, \dots, \xi_k)$ (ismeretlen) m várható értékű és σ^2 szórásnégyzetű független normális eloszlású valószínűségi változók sorozatát. A várható értéket az $\tilde{X} = \frac{1}{k}\sum_{j=1}^{k} \xi_j$ a szórásnégyzetet pedig az $S = \frac{1}{k-1}\sum_{j=1}^{k}(X_j - \tilde{X})^2$ kifejezéssel szokás becsülni. Vizsgálni szokták ezen becslések statisztikai tulajdonságait. Vegyük észre, hogy az

$$\left(\frac{\sqrt{k}}{\sigma}(\tilde{X} - m), \frac{(k-1)}{\sigma^2}S\right)$$

vektor eloszlása megegyezik a tételben vizsgált (U, V) vektor eloszlásával. Ennek tulajdonságai adnak magyarázatot sok statisztikai módszer eredetére, speciálisan az U-statisztikák megjelenésére is.

Bizonyítás: Tekintsük az $(\bar{X}_1, \dots, \bar{X}_k, \bar{X})$ véletlen vektort. Ez nulla várható értékű normális eloszlású véletlen vektor, mert egy normális vektor lineáris transzformációja is normális eloszású. Számítsuk ki ennek a véletlen vektornak a kovariancia mátrixát. Vegyük először észre, hogy $\operatorname{Var}\bar{X} = \frac{1}{k}$, és

$$\begin{aligned}\operatorname{Cov}(\bar{X}_j, \bar{X}) &= \frac{1}{k}\left(\left(1 - \frac{1}{k}\right)EX_j^2 - \frac{1}{k}\left(EX_1^2 + \dots + EX_{j-1}^2 + EX_{j+1}^2 + \dots + EX_k^2\right)\right) \\ &= \frac{1}{k}\left(\left(1 - \frac{1}{k}\right) - \frac{k-1}{k}\right) = 0.\end{aligned}$$

Innen következik, hogy $(\bar{X}_1, \dots, \bar{X}_k)$ egy a $\bar{X}$ normális eloszlású valószínűségi változótól független nulla várható értékű normális eloszlású nulla várható értékű véletlen vektor, és $\bar{X}$ nulla várható értékű $\frac{1}{k}$ szórásnégyzetű normális eloszlású véletlen vektor.

Továbbá,

$$\operatorname{Var}\bar{X}_j = \left(1 - \frac{1}{k}\right)^2 \operatorname{Var}X_j^2 + \frac{k-1}{k^2}\operatorname{Var}X_1^2 = \frac{(k-1)^2 + (k-1)}{k^2} = \frac{k-1}{k},$$

és

$$\operatorname{Cov}(\bar{X}_j, \bar{X}_l) = -\frac{1}{k}\frac{k-1}{k}(EX_j^2 + EX_l^2) + (k-2)\frac{EX_1^2}{k^2} = -\frac{2(k-1)}{k^2} + \frac{k-2}{k^2} = -\frac{1}{k},$$

ha $j \neq l$. Ez azt jelenti, hogy alkalmazhatjuk a χ^2 próbáról szóló tétel redukcióját $p_j = \frac{1}{k}$, $1 \leq j \leq k$, választással, és az azt adja, hogy a $V = \sum_{j=1}^{k} \bar{X}_j^2$ valószínűségi változó $\chi^2(k-1)$ eloszlású. Továbbá láttuk, hogy V független a $\sqrt{k}\bar{X} = U$ valószínűségi változótól, és ez utóbbi standard normális eloszlású. A tételt beláttuk.