Eötvös Loránd Turományegyetem

Ipari katasztrófák előadás

Lineáris függvénykapcsolat vizsgálata nagyrendszerek esetén statisztikus eszközökkel

Szerző:
Rudolf Ádám -
Fizikus MSc.,
II. évfolyam

Előadó:
Dr. Makai Mihály -
BME NTI

2013. június 10.

1. Bevezetés

Az ipari rendszerek, üzemek, bonyolult gépek általános leírására gyakran *nagyrendszereket* használunk. Ezek absztrakt rendszerek, amiknek sok számszerűsíthető paraméterük van, amik bonyolult, esetleg ismeretlen folyamatok eredményeként állnak elő. Ezeket modellezni tudjuk valamilyen eloszlású véletlen változókkal. A rendszernek van tehát adott számú paramétere, amik az idő előrehaladtával véletlenszerűen, de megadott (ismert, vagy ismeretlen) eloszlás szerint változnak. Ennek szimulálására programokat lehet írni, majd szimulálni a rendszer működését, és ezzel vizsgálni a paraméterek esetleges extrém értékeit, összefüggéseit.

Feladatom a fiktív ELTEc nevű nagyrendszer vizsgálata volt. Ennek $N_c = 100$ paramétere van, ezek eloszlása normális. Az i-ik paraméter várhatóértéke p_i, szórása s_i. A rendszer akkor tekinthető biztonságosnak, ha az O_1, O_2, O_3 paraméterek alatta maradnak az M_1, M_2, M_3 limiteknek. A kérdés az, hogy hogyan állapítható meg, hogy fennáll-e az O_1, O_2, O_3 paraméterek között lineáris függvénykapcsolat?

Az ilyen rendszerek leírására és vizsgálatára a statisztika széles és jól definiált eszköztárral rendelkezik.

2. Modellalkotás

A nagyrendszerek paramétereinek egy bizonyos kombinációját a rendszer állapotának nevezzük. Jelölje ezeket $O_i(t)$, ahol $i = 1, \ldots, N_c$. Ezeket diszkrét időpontokban szokás vizsgálni: $O_i(t_k)$, ahol $i = 1, \ldots, N_c$ és $t_0 = 0 < t_1 < t_2 < \ldots < t_n$. A jelölés egyszerűsítése érdekében azonban vezessük be a $t_k \longrightarrow k$ helyettesítést. A szimulációkat ismételten lefuttatva ugyanarra az értékre több eredményt is kapunk. Tehát ha a szimulációt M-szer futtatjuk, a következő minta áll rendelkezésünkre:

$$O_{i,j}(k), \quad \text{ahol} \quad (i = 1, \ldots, N_c), \quad (k = 1, \ldots, n), \quad (j = 1, \ldots, M) \tag{1}$$

2.1. Alapmennyiségek

Nagy M esetén becsülni tudjuk a statisztikai paramétereinket. A p_i átlag becslése:

$$\tilde{p}_i(k) = \frac{1}{M} \sum_{j=1}^{M} O_{i,j}(k) \tag{2}$$

Az s_i^2 szórásnégyzet becslése:

$$\tilde{s}_i^2(k) = \frac{1}{M} \sum_{j=1}^{M} (O_{i,j}(k) - \tilde{p}_i(k)) \tag{3}$$

$M \longrightarrow \infty$ esetben ezek elvileg konvergálnak a becsült értékekhez.

3. Korrelációk bevezetése

Feladatunk a paraméterek közötti összefüggések keresése, szükségünk van tehát egy olyan mennyiségre, ami azt jellemzi, hogy két változó mennyire változik együtt, azaz mennyire *korrelál*. Egy ξ és egy η valószínűségei változó korrelációja alatt a következőt értjük:

$$\mathcal{C}\{\xi|\eta\} = \frac{\mathbf{E}\{(\xi - \mathbf{E}\{\xi\})(\eta - \mathbf{E}\{\eta\}\}}{\mathbf{D}\{\xi\}\mathbf{D}\{\eta\}}, \tag{4}$$

ahol $\mathbf{E}$ a várható érték, $\mathbf{D}$ pedig a szórás képzés operátora.

3.1. Pearson-féle korreláció

A korreláció becslésének legelterjedtebb formája a Pearson-féle korrelációs együttható, ami a következőféleképpen áll elő:

$$r_{l,i}(k) = \frac{1}{M} \sum_{j=1}^{M} \frac{[O_{i,j}(k) - \tilde{p}_i(k)]\,[O_{l,j}(k) - \tilde{p}_l(k)]}{\tilde{s}_i(k)\tilde{s}_l(k)} \tag{5}$$

Ez a statisztikai függvény a $\mathcal{C}_{l,i}(k)$ korrelációs együttható becslésére szolgál. Ez a szám a változók közötti lineáris kapcsolatra érzékeny, ezt kell tehát használni esetünkben. Bizonyítás nélkül: egy ξ és egy η valószínűségi változókra $\mathcal{C}\{\xi|\eta\} = 1$ akkor és csak akkor, ha ξ és η között lineáris függvénykapcsolat van. Az $r_{l,i}(k)$ mennyiség azonban csak empirikus becslést ad a valódi korrelációról, vagyis valahogy el kell döntenünk, hogy adott hiba mellett valódi effektust jelez a kapott érték, vagy pedig csak a mintából származó ingadozás jelentkezik.

4. Hipotézisvizsgálat

Megmutatható, hogy Adott $\mathcal{C}\{\xi|\eta\}$ korrelációs együttható esetén a becslésre adott $r_{l,i}(k)$ valószínűségi változó eloszlása, ha η és ξ normális eloszlásúak:

$$f_M(r,\mathcal{C}) = \frac{M-2}{\pi}(1-\mathcal{C}^2)^{(M-1)/2}(1-r^2)^{(M-4)/2}h_M(r\mathcal{C}), \tag{6}$$

ahol

$$h_M(r\mathcal{C}) = \int_0^1 \frac{x^{M-2}}{\sqrt{1-x^2}(1-r\mathcal{C}x)^{M-1}}dx. \tag{7}$$

Látható a (6) formulából, hogy $\mathcal{C} = 1$ esetén az eloszlás a $(1-\mathcal{C}^2)$ tag miatt majdnem mindenütt 0-hoz tart. Ez azt jelenti, hogy ha a $\mathcal{C} = 1$ egzaktul teljesül, akkor r becslésre is pontosan 1-et fogunk kapni. Ellenkező esetben adott α konfidenciaszint mellett konfidencia-intervallummal fedhetjük le $\mathcal{C}$ valósi értékét. Ez azt jelenti, hogy pl. $\alpha = 5\%$-os konfidenciaszint és adott r becslés mellett megadjuk azt az intervallumot, amibe $\mathcal{C}$ valódi értéke $1-\alpha = 95\%$ valószínűséggel beleesik. Ezzel azért kell foglalkoznunk, mert $\mathcal{C}$ nem egzakt 1 értéke esetén is lehet lineáris függvénykapcsolat egyrészt a mérési hibák, másrészt az adatok esetleges transzformációja miatt.

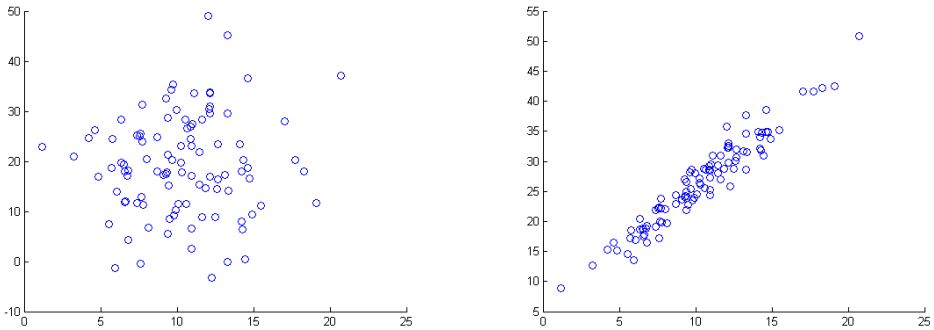
r eloszlása bonyolult és aszimmetrikus, így nehéz lenne belőle meghatározni a konfidencia-intervallumot. Az eloszlás szimmetrizálása érdekében bevezetjük a

$$z = \frac{1}{2}\ln\frac{1+r}{1-r} \tag{8}$$

mennyiséget. A valódi $\mathcal{C}$ értékek is ugyanilyen módon transzformálódnak. Az így kapott z eloszlását már közelíthetjük $\mathcal{Z} = \frac{1}{2}\ln\frac{1+\mathcal{C}}{1-\mathcal{C}}$ várható értékű és $\mathbf{D}^2 = \frac{1}{M-3}$ szóórásnégyzetű normálissal, amiből már ki tudjuk számítani a konfidencia-intervallum határait, amit majd visszatranszformálunk az r-nek megfelelő értékekre. Ha megkapjuk a z értéket, akkor a konfidencia-intervallum határai $z_\pm = z \pm \mathbf{D}\gamma$ lesznek, ahol γ a konfidenciaszinttől függő érték. Ez $\alpha = 0,05$ esetén $\gamma = 1,96$. Ezeket utána az $r_\pm = \operatorname{th} z_\pm$ inverz művelettel visszatranszformáljuk. Ezek lesznek a keresett konfidencia-intervallum határai.

5. Példa

A fent leírtak konkrét adatokon történő bemutatására MATLAB programot írtam. Ezzel $M = 100$ futásra legeneráltam az O_1, O_2 és O_3 adatsorokat a következőképpen. Az O_1-hez $p_1 = 10$, $s_1 = 3$-t választottam, O_2-höz pedig $p_2 = 20$, $s_2 = 10$-t. Ezek egymástól független, azaz korrelálatlan adatok lesznek. O_3-at a következőképpen generáltam le: $O_{3,j} = 2 \cdot O_{1,j} + 0,2 \cdot O_{2,j} + 2 \quad \forall j$-re. Az időfüggéssel nem foglalkoztam, mindezt egy időpontra számítottam ki.



(a) O_1-O_2 szórási diagram (b) O_1-O_3 szórási diagram

1. ábra. Látszik, hogy az O_1 és O_2 változók függetlenek, az O_1 és O_3 változók között viszont hibával terhelve ugyan, de erős korreláció van.

A MATLAB beépített `corr` függvényével kiszámoltam az adatsoraim közötti korrelációkat. Most az O_1-O_2 és az O_1-O_3 adatokkal fogok foglalkozni. Ezekre az $r_{1,2} = 0,0754$ és $r_{1,3} = 0,9626$ értékeket kaptam. Ezeknek a $z_{1,2} = 0,076$ és $z_{1,3} = 1,980$ értékek felelnek meg. 5 %-os konfidenciaszinten mindkettőhöz a $\pm\mathbf{D}\gamma = \frac{1,96}{\sqrt{97}} = 0,199$ értéket kell hozzáadni a határok kiszámításához. Így a $z_{1,2,-} = -0,123$, $z_{1,2,+} = 0,275$, valamint a $z_{1,3,-} = 1,781$ és $z_{1,3,+} = 2,179$ értékeket kapjuk. A visszatranszformálás után:

r	**-**	+
1,2	-0,123	0,268
1,3	0,945	0,975

1. táblázat. Korrelációs együtthatók konfidencia-intervallumának határai. Látszik, hogy az 1 és a 3 változó között jóval nagyobb értékeket kapunk, mint az elvileg független 1 és 2 adatsorok között.

Az O_1 és O_3 értékek szándékosan úgy voltak megkonstruálva, hogy legyen közöttük valamilyen, 1-nél kisebb, de hozzá közeli korreláció. Ezzel összhangban vannak a kapott eredmények. Az O_1 és O_2 adatsorok elvileg függetlenek, így 0-hoz közeli értéket várunk. A rájuk kapott r érték valóban kicsi, és a konfidencia-intervallum tartalmazza a 0-t, ráadásul közel szimmetrikus rá. Hogy az egzakt lineáris kapcsolaton kívül mit nevezünk lineáris függvénykapcsolatnak a fizikában, az rajtunk múlik, ám az 1-hez közeli korrelációs együtthatóval rendelkező adatpár halmazokat általában lineárisnak tekinthetjük, vagyis kijelenthetjük, hogy az O_1 és O_3 adatsorok lineáris kapcsolatban vannak, ám valamelyik adatsor (vagy mindkettő) hibával terhelt. Az adatsorok konstruálásánál pontosan így próbáltam létrehozni az értékeket.

6. Összefoglalás

Dolgozatomban a matematikai háttér részletezése nélkül megmutattam, hogyan kell normális eloszlású valószínűségi változók között lineáris kapcsolatot keresni. Bemutattam a felhasznált matematikai mennyiségeket, az idevágó módszereket az adatok analíziséhez. Példa gyanánt generáltam adott normális eloszlású független és korrelált adatsorokat, majd elvégeztem rajtuk a leírt analízist, ami a várt eredményt hozta.

Felhasznált irodalom:

- PÁL Lénárd: Megjegyzések nagy rendszerek bizonytalansági analíziséhez