Az AMC technikai hírlevelek informális, de irányadó közlönyök az analitikai társadalom számára érdekes technikai ügyekről. Az RSC Analitikai Részlegének Analitikai Módszerek Bizottsága adja ki, gondosan lektorálva.

Levelezési cím: The Anaytical Methods Committee, The Royal Society of Chemistry, Burlington House, Piccadilly, London W1V 0BN.

A technikai hírlevelek a webhelyen megtalálhatók: http://www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/TechnicalBriefs.asp

Robusztus statisztika: módszer a kiugró értékek kezelésére

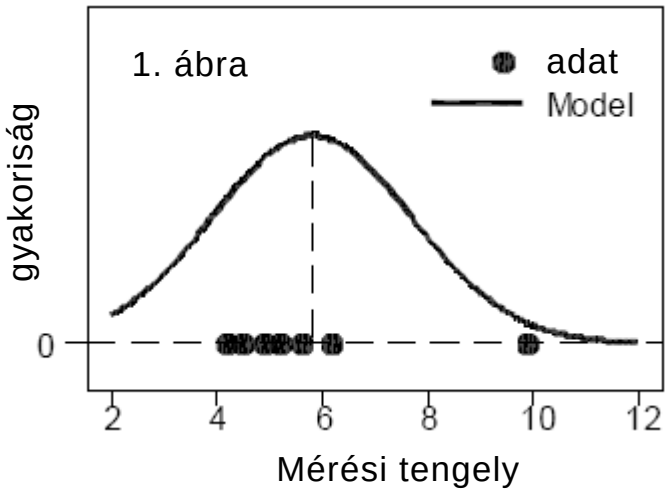
A robusztus statisztika kényelmes, korszerű módja az eredmények összegzésének, ha azt gyanítjuk, hogy kis arányban kiugró értékek is vannak köztük. A centrális tendencia legtöbb becslése (pl. számtani közép) és a szóródás (pl. szórás) értelmezése azon az implicit feltételezésen alapul, hogy az adatok egy normál eloszlásból vett véletlenszerű mintát alkotnak. Tudjuk viszont, hogy az analitikai adatok gyakran eltérnek ettől a modelltől. Gyakran erősen torzultak (a vártnál nagyobb arányban tartalmaznak az átlagtól távolabbi eredményeket), időnként pedig kiugró értékek vannak köztük.

Nézzünk példaként egy adatsorozatot:

4,5 4,9 5,6 6,2 5,2 9,9

A 9,9-es érték egyértelműen gyanús, még egy ilyen kis minta esetén is. Ha a gyanús értéket is bevonjuk a statisztikába, a következőt kapjuk:

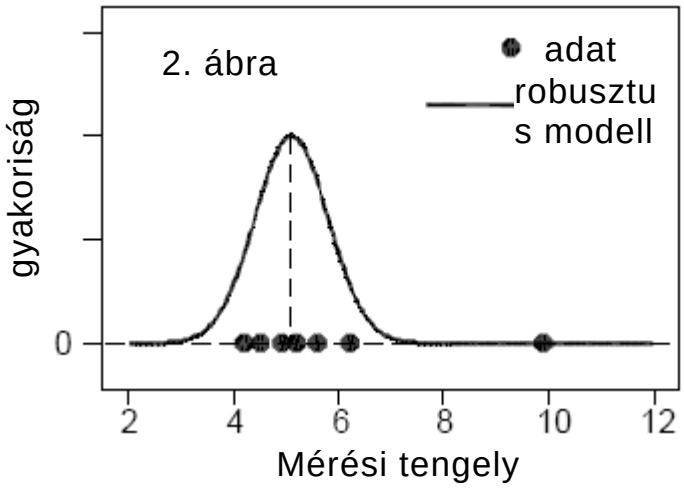
$\bar{x} = 5{,}8 \qquad s = 1{,}9$



Ezek a statisztikák normál eloszlású modell alapján leírják az adatokat, de nem jól (1. ábra). Az átlagnak nagy a torzítása, míg a szórás túl nagy. Ezenkívül e becslések számértékei, különösen a szórás, erősen függenek a gyanús adat tényleges értékétől.

Ezeknek az adatoknak ésszerűbb értelmezése, hogy azok a populáció véletlenszerű mintáját alkotják, átlaguk 5, a szórás kb. 1, és van egy kiugró érték, a 9,9. Ha a kiugró értéket kihagyjuk a számításunkból, akkor azt kapjuk, hogy:

$\bar{x} = 5{,}1 \qquad s = 0{,}7$



Ez a statisztika a legtöbb adat esetében plauzibilis normál modellt szolgáltat (2. ábra). Bár nem figyelmeztet minket a kieső adat esetleges jelenlétére, ez a modell sokszor kedvezőbb az analitikai alkalmazásokban.

A kiugró értékek vizsgálata és robusztus módszerek

A gyanús értékeket tipikusan olyan vizsgálatokkal kezeljük, mint pl. a Dixon vagy Grubb teszt a kiugró értékek meghatározására, meghatározott konfidencia szinten. Ez az eljárás nem szükségszerűen egyértelmű. Először is, a próbák egyszerű változatai félrevezethetnek, ha két vagy több kiugró érték is jelen van. Másodszor, el kell döntenünk, kihagyjuk-e a kiugró értékeket a további statisztikából. Ez felveti azt a vitatott kérdést, hogy mikor jogos a kiugró értékek elhagyása.

A robusztus statisztika alternatív eljárást kínál, mely modellt ad az adatok „jó" részének leírására, de nem követeli meg tőlünk, hogy bizonyos észlelési pontokat kiugró értékként határozzunk meg vagy elhagyjuk őket. Az átlag és a szórás becslésére sok különböző robusztus becslés létezik. Először egy nagyon egyszerű módszert tekintünk át, majd egy bonyolultabbat.

A medián/MAD módszer

Ebben a módszerben egyszerűen a rendezett adatok centrális értékét (a mediánt) vesszük az átlag becsléseként.

4,2 4,5 **5,2** 5,6 6,2 9,9

Megjegyezzük, hogy a medián nem változik, akármennyivel is növeljük a kiugró értéket. A medián az átlag robusztus becslése, melyet a $\hat{\mu} = 5{,}2$ ad meg. A becslést $\hat{\mu}$ -nek (ejtsd kalapos mu, azaz angolul mu-hat) nevezzük, hogy megkülönböztessük a szokásos számtani középtől, $\bar{x}$ -tól.

A szórás becsléséhez először kiszámítjuk az egyes értékek és a medián közti különbséget, ami ugyanabban a sorrendben a következő:

-1,0 -0,7 -0,3 0,0 0,4 1,0 4,7

Ezután nagyság szerinti sorrendbe rendezzük a különbségeket (azaz az előjelre való tekintet nélkül), és ezeknek az értékeknek a mediánját állapítjuk meg (a medián abszolút különbségét, amit MAD-del jelölünk)

Ez a következő:

0,0 0,3 0,4 **0,7** 1,0 1,0 4,7

A MAD értéke tehát 0,7. Ismét megjegyezzük, hogy a kiugró érték növelése nem befolyásolja MAD értékét. A robusztus szórás becslését úgy állapítjuk meg, hogy a MAD-et egy 1,5-höz közeli értékkel szorozzuk be. Ez adja meg a „kalapos szigma" robusztus értéket, $\hat{\sigma} = 1{,}05$.

A Huber módszer

A Huber módszer jobban kihasználja az adatok nyújtotta információt. E szerint a módszer szerint progresszíven transzformáljuk az eredeti adatokat egy winsorizálásnak nevezett eljárással [1]. Tételezzük fel, hogy a kiindulási becsléseink $\hat{\mu}_0$ és $\hat{\sigma}_0$. (Ezeket úgy értékelhetjük, mint medián-MAD becslések, vagy egyszerűen $\bar{x}$ és s). Ha egy x_i érték a $\hat{\mu}_0 + 1{,}5\hat{\sigma}_0$ –nél nagyobb, akkor megváltoztatjuk $\tilde{x}_i = \hat{\mu}_0 + 1{,}5\hat{\sigma}_0$-ra. Ugyanilyen módon, ha az érték $\hat{\mu}_0 - 1{,}5\hat{\sigma}_0$-nál kisebb, megváltoztatjuk $\tilde{x}_i = \hat{\mu}_0 - 1{,}5\hat{\sigma}_0$-ra. Egyébként pedig $\tilde{x}_i = x_i$. Ezután kiszámítjuk az átlag javított becslését , $\hat{\mu}_i$ =átlag($\tilde{x}_i$) és a szórást mint $\hat{\sigma}_i = 1{,}134$ x stdev($\tilde{x}_i$). (Az 1,134 szorzótényező a normál eloszlásból jön, a winsorizálási eljárásban leggyakrabban használt az 1,5 érték).

A példaként használt adatsorunk kissé kevés a winsorizáláshoz, de a módszert illusztrálhatjuk vele. Ha $\hat{\mu}_0 = 5{,}2$-t és $\hat{\sigma}_0 = 1{,}05$-t használjuk, a winsorizálással az adatsort a következőképpen alakítjuk át:

4,5 4,9 5,6 4,2 6,2 5,2 **6,775**

a javított becslések pedig $\hat{\mu}_1 = 5{,}36$ és $\hat{\sigma}1_0 = 1{,}15$. Az eljárás lassan konvergál, így a módszer kézi számolásra nem alkalmas. Az algoritmus Minitab-os kivitelezése az AMC szoftverben található.

Egyéb robusztus statisztikák

Összetettebb típusú statisztikák, például a varianciaanalízis [2] vagy a regresszió [3] is robusztussá alakítható. A robusztus varianciaanalízis különösen hasznos az analitikában a körvizsgálatokból származó adatok értelmezésére [4]. A robusztus regresszió hasznos lehet a kalibrálásban, de erre még nincsenek analitikai vizsgálatok.

Figyelmeztető megjegyzés

Az átlag és a szórás robusztus becslésének alkalmazása a normál eloszlás jövőbeli értékeinek előrejelzésére félrevezetheti az elővigyázatlant, mivel a kiugró értékek jelenlétére vagy valószínűségére nem ad becslést. A konfidencia-intervallumok meghatározására a robusztus becslést alkalmazni sokszor hasznos, de a kapott értékeket csak javaslatnak kell tekinteni, és nem pontos értelmezésnek.

Mikor ne használjunk robusztus módszereket

A robusztus módszerek feltételezik, hogy a vizsgált eloszlás durván normális (ezért unimodális és szimmetrikus), de kiugró adatokkal és erős torzulással (tails) szennyezett. A módszerek félrevezető eredményt adhatnak, ha olyan adatsorokra alkalmazzuk, melyek jelentősen ferdültek vagy multimodálisak, illetve ha az adatok nagy része azonos értékű.

Utoljára a kiugró értékekről

Egy adatsorra robusztus statisztikai modellt felállítani valószínűleg a legjobb módszere a gyanús értékek azonosításának, további vizsgálat céljából. A példa adatainkat véve egyszerűen transzformáljuk őket $z=(x-\hat{\mu})/\hat{\sigma}$)-al. Ha $\hat{\mu}=5{,}36$-ot és $\hat{\sigma}=$ 1,15-öt alkalmazzuk, a következő eredményt kapjuk:

$$Z=[-0{,}7 \quad -0{,}4 \quad 0{,}2 \quad -1{,}0 \quad 0{,}7 \quad -0{,}1 \quad \mathbf{3{,}9}]$$

Minden 2,5-nél nagyobb értéket gyanúsnak kell tekintenünk és a kiugró értékünk világosan látható.

Hivatkozás:

[1] AMC, Analyst, 1989, **114**, 1489

[2] AMC, Analyst, 1989, **114**, 1693.

[3] P J Rousseeuw, J. Chemomet, 1991, **5**, 1.

[4] P,J. Lowthian, M. Thompson, R Wood, Analyst, 1998, **123**, 2803