BUDAPESTI MŰSZAKI ÉS GAZDASÁGTUDOMÁNYI EGYETEM GÉPÉSZMÉRNÖKI KAR

MűGazd 2ZH Elmélet

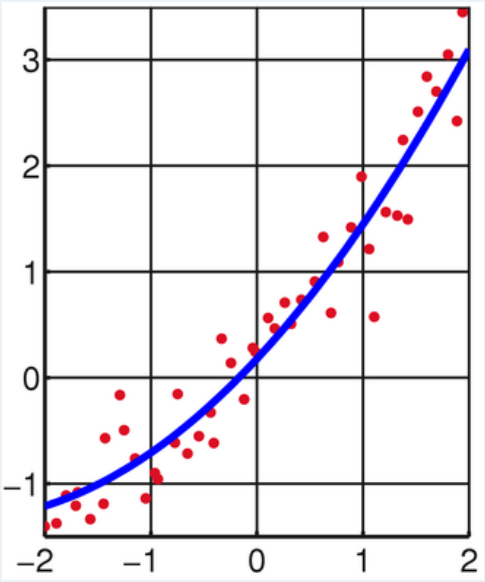
Műszaki és gazdasági adatok elemzése
(BMEGEVGBM14)

Készítette:
Kis Erhard

BUDAPEST, 2023

Legkisebb negyzetek módszere

A legkisebb négyzetek módszere esetén az adatoktól elvárjuk, hogy megfeleljenek bizonyos tulajdonságoknak, illetve éppen ellenkezőleg, hogy bizonyos tulajdonságok ne lépjenek fel. Ilyen nemkívánatos tulajdonságok a kívülálló adatok, és a multikollinearitás.
A módszer érzékeny a nagyon kilógó adatokra. Egy kilógó adat az egész eljárás eredményét megváltoztathatja, hamis képet adva az adatsorról. Különböző statisztikai tesztekkel szűrik az adatsort, hogy ne maradjanak benne mérési hibák. A kilógó adatokat elhagyják, vagy a kívülállókra kevésbé érzékeny módszerekkel alternatív becsléseket végeznek. ilyen például a súlyozott regresszió, amiben a kívülálló adatok súlyát, és ezzel befolyását is csökkentik.



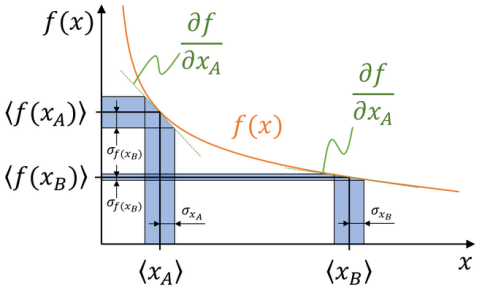
Több független változó esetén a multikollinearitás azt jelenti, hogy két független változó erősen korrelál, ezért közel állnak a lineáris összefüggéshez. Ez azért baj, mert így a feladat rosszul kondicionálttá válik, ami azt jelenti, hogy érzékeny lesz a mérési hibákra; kis hibák is nagyon eltérő eredményhez vezetnek.

Wald módszer

A statisztikában Ward módszere egy kritérium, amit hierarchikus csoportanalízis alkalmazására használnak. Ward minimum variancia módszere Joe H. Ward, Jr. által javasolt objektív függvény alapján történő hierarchikus klaszterezési eljárás egy speciális esete. Ward egy általános agglomeratív hierarchikus klaszterezési eljárást javasolt, ahol a kiválasztott klaszterek párosa minden lépésben az objektív függvény optimális értékén alapul. Az objektív függvény "bármilyen olyan függvény lehet, ami tükrözi a kutató célját". Ennek a nagyon általános osztálynak számos szokásos klaszterezési eljárás is része. A módszer bemutatásához Ward az együttes négyzetek hibáját alkalmazta, és ez az példa ismert Ward módszerként vagy pontosabban Ward minimum variancia módszerként.
A legközelebbi szomszéd lánc algoritmust használhatjuk ugyanezen klaszterezés megtalálására, időarányosan a bemeneti távolsági mátrix méretéhez és lineáris térigény mellett a klaszterezendő pontok számához képest

Hibaterjedés

A statisztikában hibaterjedésnek nevezik a származtatott mennyiségek hibájának az alapul szolgáló mennyiségek hibájától való függését, illetve magát a matematikai módszert, mellyel a származtatott mennyiségek hibáját becslik. A hibaterjedés figyelembe vétele a fizikában is gyakran használatos, ha például hibával terhelt mért mennyiségekből valamilyen összefüggés segítségével származtatott új mennyiség hibáját határozzák meg.



Például ha egy, az Ohm-törvénynek engedelmeskedő áramköri rendszeren mérjük az I átfolyó áramot, és annak ΔI bizonytalanságát, továbbá az első U feszültséget, és annak ΔU bizonytalanságát, akkor az ellenállás meghatározására szolgáló $R = f(U, I) = \frac{U}{I}$ összefüggés és a hibaterjedés figyelembe vételével a származtatott ellenállás ΔR bizonytalansága jól közelíthető. Egyes esetekben, például $f = AB$ alakú összefüggés esetén f hibája egzakt módon is kifejezhető, de általában sorfejtésen alapuló, lineáris közelítést alkalmaznak.
A hibaterjedés jellegét alapvetően az alábbiak határozzák meg:

- A kiinduló mennyiségek bizonytalanságának összefüggése illetve függetlensége befolyásolja a származtatott mennyiség hibájának számolását.
- A származtatott mennyiség kifejezését megadó f összefüggés jellege befolyásolja, hogy mely mért mennyiségek hibája milyen mértékben járul hozzá a származtatott hibához.

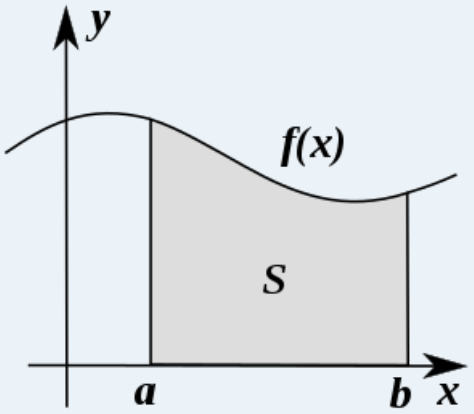
Konfidencia intervallum

A konfidenciaintervallum a valószínűségi intervallum, az induktív statisztika eszköze: ha mintából becsülünk, sosem tudjuk a pontos értéket, a teljes sokaság felmérése igen drága dolog. A konfidenciaintervallum adott szignifikanciaszinten: a becsült változó alsó és felső korlátja.
A konfidenciaintervallum intervallum értékű becslést ad egy paraméterre: valószínűleg ezek közé a korlátok közé esik. Ez sok esetben jobb, mint egyetlen becsült értéket adni. Az α paraméter egy előzetesen megadott értékére a becsült paraméter $1 - \alpha$ valószínűséggel esik az intervallumba. Ezt az $1 - \alpha$ szintet sokszor százalékban adják meg; például 95% tipikus.
Ahhoz, hogy megbízhatósági intervallumot számoljunk, különböző feltevésekkel kell élnünk, például a becslés hibáinak eloszlása normális. Ezzel a feltevéssel a megbízhatósági intervallum a következőképpen számítható:
$\mathrm{Int}(t) = t \pm (z * sh)$, ahol t vagy ϑ a sokasági paraméter, z a megadott szint (α) a normális eloszlás $1 - \alpha - d$ rendű kvantilise, sh pedig a standard hiba, ami általában a szórás osztva a minta elemszámának négyzetgyökével.

Eloszlás és Sűrűségfüggvény

A valószínűségszámításban az X valószínűségi változó sűrűségfüggvénye f pontosan akkor, ha az X-nek az F-fel jelölt eloszlásfüggvénye előállítható a következő alakban:

$$F(x) = \int_{-\infty}^{x} f(t)\ dt$$

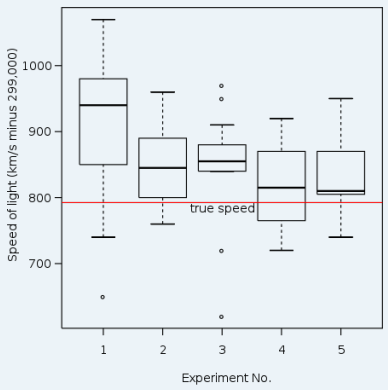


Szemben a valószínűségekkel, a sűrűségfüggvények felvehetnek 1-nél nagyobb értéket is. A valószínűségi eloszlások sűrűségfüggvényeken alapuló konstrukciója szempontjából nem a sűrűségfüggvény által felvett érték a fontos, hanem az integrál.

A sűrűségfüggvény általánosítása az általánosított sűrűségfüggvény, ahol is a Lebesgue-mértékre vonatkozó sűrűségfüggvények a valószínűségi sűrűségfüggvények. A továbbiakban sűrűségfüggvényen valószínűségi sűrűségfüggvényt értünk, kivéve ha azt máshogy jelezzük.
Diszkrét esetben az események valószínűsége megkapható a tartalmazott elemi események valószínűségeinek összegzésével. Folytonos esetben azonban ez nem tehető meg, mivel a nullaszor végtelen értéke bármi lehet. Például két ember csak ritkán pont egyforma magas, eltér egymástól egy hajszállal vagy csak néhány atomnyival. A sűrűségfüggvénnyel tetszőleges intervallum valószínűsége meghatározható, így a nullaszor végtelen probléma megkerülhető.

Box plot

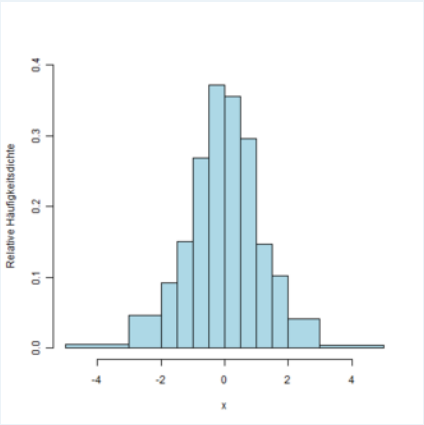
A leíró statisztikában a boxplot vagy boxplot egy módszer arra, hogy grafikusan szemléltessük a numerikus adatok csoportjainak elhelyezkedését, terjedelmét és ferdeségét a kvartileken keresztül. A boxploton kívül lehetnek vonalak (melyeket száraknak neveznek), amelyek a dobozból kiindulva mutatják a terjedelmet a felső és alsó kvartilák túlmutató változékonyságát, így a diagramot gyakran doboz-szarv diagramnak is hívják. Az adatoktól jelentősen eltérő kiugró értékeket egyedi pontként is ábrázolhatjuk a boxploton túlmutató szárakon kívül.



A boxplotok nem parametrikusak: statisztikai populáció mintaelemeinek változékonyságát mutatják be, anélkül, hogy feltételezéseket tenne az alapvető statisztikai eloszlásról (bár a Tukey-dooszdiagram a szimmetria és a normális eloszlás feltételezésével dolgozik a szárak esetében). A boxplot részszakaszainak távolságai jelzik az adatok szóródásának és ferdeségének fokát, amelyeket általában az öt számjegyű összefoglalással írnak le. Ezenkívül a boxplot lehetővé teszi különféle L-értékek vizuális becslését, különösen az interkvartilis tartomány, középferde, tartomány, középtartomány és trimean. A boxplotokat lehet vízszintesen vagy függőlegesen is rajzolni.

Hisztogram

A hisztogram metrikusan skálázott tulajdonságok grafikus ábrázolása. Ha túl sok érték szerepel, akkor osztályokba vonják össze őket. Az egyes osztályok szélessége változhat. A mennyiségeket a szorosan egymás mellé rajzolt téglalapok jelölik, ahol az egyes téglalapok területe az adott osztály gyakoriságát mutatja. A téglalapok magassága az osztály gyakorisági sűrűségét jelöli, ami az adott osztály szélességével leosztott gyakoriság.
A hisztogramok felfoghatók a folytonos valószínűségi változó sűrűségfüggvényének becsléseként.



Átlag

Számtani vagy aritmetikai középértéken n darab szám átlagát, azaz a számok összegének n-ed részét értjük. A számtani közepet általában A betűvel jelöljük:

$$A(a_1; ...; a_n) = \frac{a_1 + ... + a_n}{n}$$

Szórás

A szórás a valószínűségszámításban az eloszlásokat jellemző szóródási mérőszám. A szórás egy valószínűségi változó értékeinek a várható értéktől való eltérésének a mértéke.

Medián

A medián a statisztika egy nevezetes középértéke, úgynevezett helyzeti középérték: az az érték, amelytől mérve az elemek abszolút távolságainak összege minimális. Meghatározása: véges elemszámú sokaság esetén a medián a sorba rendezett adatok közül a középső érték, vagy másképpen: a medián az az érték, amely a sorba rendezett adatokat két egyenlő részre osztja. A gyakorlatban problémát jelent, ha páros számú adat vagy ismétlődő értékek vannak. Folytonos valószínűségi változó esetén a mediánnál húzott függőleges vonal a valószínűségsűrűségi függvény görbe alatti területét pont elfelezi.

Korreláció

A matematikában (a statisztikában) a korreláció jelzi két tetszőleges érték közötti lineáris kapcsolat nagyságát és irányát (avagy ezek egymáshoz való viszonyát). Az általános statisztikai használat során a korreláció jelzi azt, hogy két tetszőleges érték nem független egymástól. Az ilyen széles körű használat során számos együttható, érték jellemzi a korrelációt, alkalmazkodva az adatok fajtájához.
A korreláció csak a lineáris kapcsolatot jelzi. Például egy valószínűségi változó és négyzete korrelációja lehet nulla. Ha két véletlen mennyiség korrelációja nulla, akkor korrelálatlanok; ilyenkor a kapcsolatot, ha van, másként kell jellemezni, például feltételes valószínűségekkel. A normális eloszlású valószínűségi változókra jellemző, hogy ha korrelálatlanok, akkor függetlenek is. Így a korreláció jól alkalmazható normális eloszlásúnak tekinthető mérhető mennyiségek közötti kapcsolat erősségének mérésére.

Tapasztalati szórás és Korrigált tapasztalati szorás

A mintatapasztalatiszórása azt mutatja meg, hogy az egyes eredmények átlagosan mennyire térnek el az átlagtól, azaz mennyire szóródnak. A tapasztalati szórás kiszámításánál meghatározzuk az egyes értékek és az átlag különbségét, ezeket az eltérések négyzetre emeljük és összeadjuk, az összeget elosztjuk a darabszámmal és az egészből négyzetgyököt vonunk.
Ahol n a mérések darabszáma. A tapasztalati szórás általában eltér az elméleti szórástól. Amennyiben viszont korrigált tapasztalati szórást számolunk, azaz nem n-nel, hanem $(n-1)$-gyel osztunk, akkor a tapasztalati szórás várható értéke egyenlő lesz az elméleti szórással.

Tapasztalati szórás:

$$s_N = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Korrigált tapasztalati szorás:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$