STATOKOS JEGYZET: STATISZTIKA

StatOkos – Statisztikai és Módszertani Adatbázis

2018

Tartalom

[LEÍRÓ STATISZTIKA]

szerzők: Kazinczi Csaba, Holczer Adrienn, Alter Emese

Tárgymutató

Skálatípusok, mérési szintek

Nominális változók

- névleges, minőségi különbség az egyes kategóriák között
- a különbség az adatok között nem mérhető, a különböző kategóriák nem rendezhetők sorrendbe
- a két kategórás nominális változókat (pl.: nem) dichotóm változóknak nevezzük
- ilyen változó például: nem (nő/férfi), egyetemi szak, stb.
- elemzéskor használható eljárások: leíró statisztikai eljárások, pl.: eloszlás vagy a leggyakoribb érték (módusz) vizsgálata

Ordinális változók

- a különbség inkább minőségi, mint mennyiségi, nem mérhető pontosan
- az egyes kategóriák sorba rendezhetők
- ilyen változó például: jövedelemkategóriák, iskolai végzettség, versenyen elért helyezés
- használható középérték mutatók: medián, módusz

Metrikus változók

- mennyiségi különbség az egyes értékek között
- az értékek közti távolság mérhető, nem változik
- matematikai műveletek elvégzésére is alkalmas
- 2 altípus: intervallum-és arányskála
- intervallum skála: nincs természetes 0 pont
- arányskála: rendelkezik természetes 0 ponttal
- középérték mutatók: átlag, medián és módusz egyaránt használható

Középérték mutatók

Átlag

- csak metrikus változók esetében alkalmazható
- kiugró értékek könnyen torzíthatják, ilyenkor érdemes mediánt használni
- képlet

Medián

- középső érték
- ordinális és metrikus változók esetén alkalmazható

- ha az átlag az eloszlás miatt nem nyújtana megfelelő információt a minta tulajdonságáról, helyette használhatjuk a mediánt (egyes nemparaméteres próbák is a mediánnal számolnak átlag helyett)

Módusz

- a leggyakrabban szereplő érték
- lehet 1 érték, ekkor unimodális módusznak nevezzük
- amikor több érték is ugyanannyi alkalommal szerepel, és ezek szerepelnek a legtöbbször, több móduszunk lesz, ezt multimodális módusznak nevezzük
- nominális, ordinális és metrikus változók esetén is alkalmazható
-

Skálatípusok és középérték mutatók

A skála típusa	Középérték mutatók (használható-e)		
	Átlag	Medián	Módusz
Nominális	NEM	NEM	**IGEN**
Ordinális	NEM	**IGEN**	**IGEN**
Metrikus	**IGEN**	**IGEN**	**IGEN**

Szóródási mutatók

- a középérték körüli koncentrálódást, az ingadozás mértékét adják meg

A terjedelem

- a legnagyobb és legkisebb érték különbsége
- metrikus skálák esetében használható

A szórás

- a középértéktől való eltérés átlagát adja meg
- a középérték +- 1 szórás között található a legtöbb érték, az ezen kívül eső értékek kiugrónak minősülnek
- kiszámításának lépései: 1: átlag kiszámítása, 2: Egyes adatok kivonása az átlagból, 3: Kapott különbségek négyzetre emelése, 4: Négyzetre emelt értékek átlagának kiszámítása, 5: Gyökvonás a kiszámított átlagból

Eloszlás

Normáleloszlás

- Gauss-görbe vagy haranggörbe néven is ismert
- természetes változók esetén szokott előfordulni
- átlag +-1 szórás távolságra esik a minta 68%-a, átlag +-2 szórásra pedig a minta 95%-a
- a paraméteres próbák előfeltétele a normáleloszláshoz való közelítés
- nem normális eloszlás: ferdeség és csúcsosság

Ferdeség

- az eloszlás x tengelyen való jobbra vagy balra tolódása

- balra tolódás esetén pozitív ferdeség
- jobbra tolódás esetén negatív ferdeség
- a módusz, medián és átlag torzításával jár

Csúcsosság

- a normáleloszlástól y tengelyen való eltérés
- a normáleloszlás csúcsosságának értéke nulla
- normáleloszlásnál nagyobb csúcsosság: leptokurtikus, normáleloszlásnál kisebb mértékű csúcsosság, laposság: platykurtikus csúcsosság
- normáleloszláskor tapasztalható csúcsosság: mezokurtikus

[SZIGNIFIKANCIA PRÓBÁK]

szerzők: Kazinczi Csaba, Alter Emese

Tárgymutató

Illeszkedésvizsgálat

- Alkalmazás: Olyan esetekben, mikor meglévő eloszláshoz szeretnénk hasonlítani egy változó eloszlását
- pl.: A nemek eloszlása a mintámon 65-35% a nők javára. Ezt összevethetjük az átlagos, 50-50%-os eloszlással. Ekkor arra a kérdésre keressük a választ, hogy a nemek eloszlása a mi mintánkon különbözik-e szignifikánsan az átlagtól, amely 50-50%.
- Nullhipotézis: Az általunk vizsgált eloszlás nem tér el a meglévő, előre megadott értéktől.
- Ha az eredmény szignifikáns, akkor kimondhatjuk, hogy a két eloszlás egymástól szignifikánsan különbözik.

Khí-négyzet próba

- nemparaméteres eljárás
- két nominális változó kapcsolatának vizsgálata
- pl.: Van-e különbség a fiúk és a lányok között abban, hogy 3 szín közül (rózsaszín, kék, sárga) melyik játékot választják ki szabad választás esetén?
- Ebben az esetben a nem (2 kategória) és a játék színe (3 kategória) a két vizsgált változó, a kérdés pedig arra vonatkozik, hogy az eloszlás eltér-e a véletlen választás során kapott eloszlástól
- Nullhipotézis: A két változó egymástól független
- Szignifikáns eredmény esetén a két változó nem független egymástól, interakciójuk befolyással van az eloszlásokra

McNemar és Cochran-Q próbák

- nemparaméteres eljárások
- nominális, dichotóm változók eloszlásának összehasonlítása
- Pl.: Ugyanolyan ízű fagylalt 2 különböző receptjének preferenciája (kedveli/nem kedveli)
- Ebben az esetben arra a kérdésre keressük a választ, hogy a két recept esetén ugyanakkora arányban jelentek-e meg a kedveléssel kapcsolatos igen és nem válaszok
- Nullhipotézis: A két változó eloszlása megegyezik
- Szignifikáns eredmény esetén kimondhatjuk, hogy a két változó eloszlása szignifikánsan különbözik

- 2 változó esetén McNemar próbát használunk, 2-nél több változó esetén érdemes Cochran-Q próbát alkalmazni

Egymintás t-próba

- paraméteres eljárás
- Alkalmazása: Metrikus, normál eloszlású változó átlagának összevetése egy másik, előre megadott átlaggal
- pl.: Kikeresünk egy szakértők által meghatározott, sok éves átlaghőmérsékletet, és ezzel összevetjük az általunk x ideig, naponta rögzített hőmérséklet átlagát
- Nullhipotézis: Az átlagok megegyeznek
- Szignifikáns eredmény esetén elmondható, hogy az átlagok közt szignifikáns különbség van

Egymintás Wilcoxon-próba

- nemparaméteres eljárás
- az egymintás t-próba nemparaméteres párja
- nemparaméteres változó mediánjának összevetése külső értékkel
- Nullhipotézis: Nincs különbség a két medián között
- Szignifikáns eredmény esetén elmondható, hogy a mediánok szignifikánsan különböznek

Páros mintás t-próba

- paraméteres eljárás
- két metrikus, normál eloszlású változó átlagának összehasonlítása
- kontroll-feltételes elrendezés, tehát ugyanazon csoporton végzett két mérés eredményeit vetjük össze
- pl.: reakcióidő összehasonlítása sörivás előtt és után
- Nullhipotézis: A két átlag megegyezik.
- Szignifikáns eredmény esetén elmondható, hogy a két feltétel esetén kapott átlagok között szignifikáns különbség van

Páros mintás Wilcoxon-próba

- nemparaméteres eljárás
- a páros mintás t-próba nemparaméteres párja
- két nemparaméteres változó mediánjainak összehasonlítása kontroll-feltételes elrendezésben
- Nullhipotézis: A két medián nem különbözik
- Szingifikáns eredmény esetén elmondható, hogy a két mérési alkalom során kapott mediánok között szignifikáns különbség van

Kétmintás t-próba

- független mintás t-próba néven is találkozhatunk vele
- paraméteres eljárás
- két metrikus, normál eloszlású változó átlagának összehasonlítása
- kontroll-csoportos elrendezés, tehát két csoport átlagát hasonlítjuk össze
- pl.: idősebb és fiatalabb csoport reakció idejének összehasonlítása

- Nullhipotézis: A két átlag nem különbözik
- Szignifikáns eredmény esetén elmondható, hogy a két csoport átlaga között szignifikáns különbség van

Mann-Whitney próba

- nemparaméteres eljárás
- A kétmintás t-próba nemparaméteres párja
- két nemparaméteres változó mediánjainak összehasonlítása kontroll-csoportos elrendezésben
- Nullhipotézis: A két medián nem különbözik
- Szingifikáns eredmény esetén elmondható, hogy a két csoport mediánjai között szignifikáns különbség van

[TÖBBVÁLTOZÓS SZIGNIFIKANCIA PRÓBÁK]

szerzők: Kazinczi Csaba, Alter Emese

Többváltozós szignifikancia próbák

- A klasszikus szignifikancia próbák és az egyszerűbb többváltozós statisztikák határán helyezkednek el.

- Szignifikáns különbséget keresnek az egyes csoportok/feltételek között, azonban kettőnél több változóval dolgoznak.

Cochran-Q próba

- 2-nél több, nominális, dichotóm változó vizsgálatára szolgál egy csoporton belül.

- pl.: Egy osztályban vizsgáljuk a barna és szőke hajúak, zöld és kékszeműek, illetve fiúk és lányok szerint az eloszlást.

- Szignifikáns eredmény esetén elmondható, hogy az eloszlások között jelentős különbség van.

- Hivatkozása: $\chi 2$(szabadságfok - df, N = elemek száma) = McNemar vagy Cochran Q próba értéke , p = szignifikancia

Friedman-teszt

- 2-nél több ordinális változó középértékeinek vizsgálatára szolgál 1 csoporton belül.
- Az ismételt méréses ANOVA nemparaméteres párja.
- Szignifikáns eredmény esetén elmondható, hogy jelentős különbség van a változók középértékei között.

Kruskal-Wallis próba

- 2-nél több csoport összehasonlítása 1 db ordinális változó tekintetében.
- Az egyszempontos ANOVA nemparaméteres párja.
- Szignifikáns eredmény esetén elmondható, hogy a csoportok között jelentős különbség van.

[TÖBBVÁLTOZÓS STATISZTIKA: VARIANCIAANALÍZIS]

szerzők: Kazinczi Csaba, Alter Emese

Tárgymutató

- Egyszempontos varianciaanalízis
- Összetartozó mintás (ismételt méréses) varianciaanalízis
- Többszempontos varianciaanalízis
- Kevert varianciaanalízis
- Kovariáns, ANCOVA

Egyszempontos varianciaanalízis

- Egy metrikus függő változó átlagát hasonlítjuk össze kettőnél több csoport között.
- Pl.: 3 korcsoport reakcióidejének összehasonlítása valamilyen feladat esetén.
- Szignifikáns eredmény esetén elmondható, hogy a csoportok átlagai között jelentős különbség van.
- A post-hoc teszt alapján tudjuk eldönteni, hogy mely csoportpárok között volt szignifikáns eltérés.
- Hivatkozása: F(a változó szabadságfoka (df) between groups, a változó szabadságfoka (df) within group) = F-érték, p = szignifikancia érték.
- Összetartozó mintás-ismételt méréses-varianciaanalízis
- Kettőnél több, ismételt mérési alkalmakat jelölő, metrikus változó átlagainak összehasonlítása egy csoporton belül.
- pl.: Reakcióidő mérése, majd összevetése ugyanazon a mintán 4; 6 és 8 óra alvás után.
- Szingifikáns eredmény esetén elmondható, hogy a mérési alkalmak során kapott átlagok között jelentős különbség van.
- Hivatkozása: F(a változó szabadságfoka (df) , a hiba szabadságfoka (Error)) = F érték, p = szignifikancia érték.

Többszempontos varianciaanalízis

- Egy metrikus mérési szintű függő változó átlagainak összehasonlítása különböző csoportok esetén 2 nominális csoportosító változó (pl.: nem és szemszín) alapján.
- A két változó függő változóra gyakorolt hatását külön-külön is vizsgálhatjuk, ezt főhatásnak nevezzük.
- A két tényező együttes kereszthatását is vizsgálni tudjuk vele, vagyis azt, hogy a két csoportosító változó interakciójának van-e szignifikáns befolyása a függő változó varianciájára.
- Hivatkozása: F(a változó szabadságfoka (df), a hiba szabadságfoka) = F-érték, p = szignifikancia érték.

A kovariáns és az ANCOVA

- A kovariánsok varianciaanalízisbe való bevonásának célja, hogy a nullhipotézis elvetésekor olyan változók hatását is figyelembe vegyük, melyek nem képezik kutatásunk részét, de befolyással lehetnek az eredményekre. Ez azt segíti elő, hogy kutatásunk kevésbé legyen idegen a természetes, a való életben előforduló helyzetektől, illetve a kontrollt is növelhetjük általa.
- A kovariáns bevonásával annak a függő változóra gyakorolt hatását leválasztjuk, és így csak az általunk vizsgált független változók hatását mérjük.
- Fontos, hogy a bevont kovariánsok korreláljanak a független változóval, de ne korreláljanak egymással.
- A kovariáns bevezetése akkor lehet indokolt, amikor a független változók által kialakított csoportok/feltételek között valamilyen különbség alapvetően fellelhető (legalábbis sejtjük), de elsősorban nem releváns a kutatás szempontjából.
- Kovariánst mindhárom, fentebb bemutatott varianciaanalízisbe bevonhatunk, a varianciaanalízis és kovariáns együttesét ANCOVÁ-nak nevezzük.
- Az ANCOVA-próbák (egyszempontos, ismételt méréses, többszempontos) értelmezése megegyezik az azonos nevű varianciaanalízisekével.

[TÖBBVÁLTOZÓS STATISZTIKA: KORRELÁCIÓ ÉS REGRESSZIÓ]

szerzők: Kazinczi Csaba, Alter Emese

A korreláció

- Változók kapcsolatának vizsgálatára alkalmas olyan esetekben, amikor nincs szilárd előfeltételezésünk a kapcsolatra vonatkozóan

- Nem alkalmas ok-okozati kapcsolat feltárására, csupán az együttjárás meglétét, erősségét és irányát tudjuk ellenőrizni vele

- Metrikus változók esetén Pearson-korrelációt, ordinális változók esetén Spearman-féle (nemparaméteres) korrelációt alkalmazhatunk

- A korrelációs együttható (r) értéke -1 és 1 közé eshet. Minél közelebb esik r abszolút értéke 1-hez, annál erősebb együttjárásról beszélhetünk. Ha r = 1, egyenes arányosságról, ha r = -1, fordított arányosságról beszélhetünk. Ha r negatív, fordított kapcsolatról beszélhetünk, vagyis minél kisebb x értéke, annál nagyobb y értéke, és fordítva. Pozitív irányú együttjárás esetén minél magasabb x értéke, annál magasabb y is, illetve minél alacsonyabb a egyik érték, annál alacsonyabb a másik is.

- Jelentős együttjárásról csak szignifikáns eredmény esetén beszélhetünk.

- Hivatkozása: r(n-elemszám) = r értéke, p = szignifikancia érték.

A lineáris regresszió

- Olyan esetekben alkalmazzuk változók kapcsolatának vizsgálatára, amikor a kapcsolatról szilárd előfeltételezésünk van.
- Nem csak a kapcsolat meglétét, de irányát is vizsgálhatjuk a segítségével.
- Annak a vizsgálatára használjuk, hogy több paraméteres változó milyen mértékben befolyásol egy szintén paraméteres függő változót. A függő és független változókat mi adjuk meg a próba lefuttatása során, tehát előfeltételezésünknek kell lennie a kapcsolat irányára vonatkozóan.
- Hivatkozása: R2 = R square értéke, F(regresszió szabadságfoka, reziduális szabadságfoka) = F értéke, p=szignifikancia értéke.

[TÖBBVÁLTOZÓS STATISZTIKA: FAKTOR-ÉS KLASZTERELEMZÉS]

szerzők: Kazinczi Csaba, Alter Emese

A faktorelemzés és a főkomponenselemzés

- A faktor-és főkomponenselemzés célja egyaránt az, hogy csökkentsük a változók számát, illetve, hogy olyan új változókat hozzunk létre, melyek egymással nincsenek szoros korrelációban (multikollinearitás kiküszöbölése).

- A két eljárás lényege, hogy segítségükkel nagyszámú változóból kisebb számú faktort hozunk létre.

- A különbség, hogy faktorelemzést akkor használunk, ha egy adott kérdőívnek ismerjük a fő tulajdonságait, hiányosságait, rendelkezünk előfeltételezéssel az egyes itemekkel és azzal kapcsolatban, hogy mely faktorba sorolandók. Abban az esetben, amikor nincsenek ilyen előzetes ismereteink és előfeltételezéseink, tehát amikor feltáró jellegű munkát végzünk, érdemes főkomponenselemzést használnunk.

A klaszterelemzés

- A faktorelemzéssel szemben a klaszterelemzés célja nem a változók számának csökkentése, hanem a minta elemeinek csoportba rendezése, ezáltal téve azt átláthatóbbá.
- A klaszterelemzés során megadott változók szerint soroljuk csoportokba a vizsgálati személyeket úgy, hogy egy-egy csoport tagjai egymáshoz hasonlóak legyenek, de más csoportok tagjaitól különbözzenek.
- Klaszterelemzés esetén a csoportokat nem definiáljuk előre, csupán a csoportok létrehozásához használt változókat adjuk meg, szemben a diszkriminanciaelemzéssel, ahol előzetes ismereteink alapján csoportosítunk.

[TÖBBVÁLTOZÓS STATISZTIKA: DISZKRIMINANCIA ANALÍZIS ÉS LOGISZTIKUS REGRESSZIÓ]

szerzők: Kazinczi Csaba, Alter Emese

A diszkriminancia analízis

- A diszkriminancia analízis a prediktív jellegű eljárások közé sorolandó.

- Fontos kiindulópontja, hogy a populációnknak vannak bizonyos állandó tulajdonságai, melyek akkor is megmaradnak, ha a tagok között bizonyos mértékű fluktuáció van.

- A diszkriminancia analízis esetében a függő változónk egy csoportosításra alkalmas változó, a független változóink pedig olyan paraméteres változók, melyek valamilyen módon jellemzik a minta tagjait, így az ezekből kinyerhető információk alapján nagy valószínűséggel helyesen be tudjuk sorolni őket a függő változó által megadott kategóriákba.

- Lehet két/többváltozós: Kétváltozós esetén a függő változó dichotóm, vagyis 2 csoporta osztja a mintát, több kategória esetén többváltozós diszkriminancia analízisről beszélhetünk.

A logisztikus regresszió

- Célja a diszkriminancia analízishez hasonlóan a minta tagjainak a kategorikus függő változó által meghatározott csoportjaiba való helyes besorolás megadott független változók alapján, ebben az esetben azonban nem csak paraméteres változók lehetnek független változók.

- 2 fő típusa: Bináris logisztikus regresszió: Ekkor a függő változó dichotóm, vagyis 2 csoportba sorolhatjuk a személyeket a független változók alapján. Többváltozós logisztikus regresszió: A függő változó 2-nél több kategóriával rendelkezik, melyekbe besorolhatjuk a személyeket a független változók alapján.