

Házi feladat II.

Papíron megoldandó feladat (természetesen a szükséges kvantiliseket lehet sőt ajánlott is Excel-ben számolni) (7 pont): Az építőmérnöki vállalkozásunk egy alvállalkozó révén jut 25 kg-os cementes zsákokhoz. A legutóbbi beszállításnál próbaképpen lemértük 10 zsák tömegét. A tömegekre a következők adódtak: 24.5, 25.1, 24.7, 24.9, 25.2, 24.8, 24.4, 24.9, 24.8, 24.7. Az alábbi hipotézisről szeretnénk $\varepsilon = 0,05$ elsőfajú hibavalószínűség mellett dönteni (korábbi tapasztalatok alapján tudjuk, hogy a zsákok tömege normális eloszlást követ):

H_0 : A cementes zsákok tömegének várható értéke legalább 25 kg

H_1 : A cementes zsákok tömegének várható értéke kisebb, mint 25 kg

Oldjuk meg a feladatot amikor

- (a) tudjuk, hogy a cementes zsákok tömegének szórása 0.2 kg;
- (b) nem ismerjük a cementes zsákok tömegének szórását.

Segítség: A t-próba egyoldali módosítása megtalálható a 10. óra táblaképei között. Figyelj oda arra, hogy a megfelelő szabadságfokú t-eloszlás kvantilisét használd az elfogadási "intervallum" kialakításakor. Az u-próba egyoldali módosítását analóg módon kell csinálni.

EXCEL feladat 1 (6 pont): A teszteredmények.xlsx file középiskolások különböző teszteken elért eredményeit tartalmazza.

- (a) Olvasd be a file-t majd ábrázold pontfelhő ábrán a tanulók math és science teszteredményeit úgy, hogy a math eredmények tengelye legyen az x tengely! A tanult módon rakd rá az ábrára az (empirikus) regressziós egyenest és annak egyenletét illetve az R^2 statisztikát!
- (b) Az (empirikus) regressziós egyenes ábrára kiírt meredekségét és y tengelymetszetét határozd meg az Excel lin.ill() függvényével is (a 3. paraméternek az IGAZ-t (!!itt a legelső verzióban elírás volt!!!), a 4. paraméternek a HAMIS-t add meg, ekkor a végeredményt úgy kapod meg, hogy egy cellába beírod a függvényt, majd kiegészítéd a kijelölést a mellette levő cellával, végül F2 majd ctrl+shift+enter megnyomása után együtt fog látszódni a meredekség és az y tengelymetszet)!
- (c) Számold ki az empirikus regressziós egyenes által a science eredményekre adódó becsléseket egy új oszlopban!
- (d) Jelölje $Y_1, Y_2, \dots, Y_n$ a lineáris regresszió célváltozójának megfigyelt értékeit, $\bar{Y}$ ezek átlagát, továbbá jelölje $\tilde{Y}_1, \tilde{Y}_2, \dots, \tilde{Y}_n$ a lineáris regresszióból adódó becsléseket. Az R^2 statisztikát az

$$1 - \frac{\sum_{i=1}^n (Y_i - \tilde{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (1)$$

formulával szokták definiálni. Számold ki az ábrára kiírt R^2 statisztikát a fenti definíció alapján is!

- (e) Kérd le a math és science változók közötti illetve a science és a becsült science változók közötti korrelációs együtthatókat majd emeld ez(eke)t négyzetre! Mit veszel észre?

EXCEL feladat 2 (7 pont): Nyisd meg a házár adatokat tartalmazó Excel file-t. Először számold ki egyszerű osztással a házak 1 négyzetlábra jutó eladási árát! Ezután vizsgáld meg azt a kérdést, hogy a négyzetlábankénti eladási ár összefügg-e az adó mértékével (TAX változó)! Mivel folytonos változókról van szó, ezért közvetlenül nem használható a chi-négyzet teszt. Ezért azt kérjük, hogy először hozzátok létre a változók diszkrétizált változatait vagyis a folytonos változók egyes értékeihez rendeljétek a "kicsi", "közepes" illetve "nagy" címkéket aszerint, hogy a megfigyelés a számegyenes következő osztópontok által meghatározott felosztásának melyik részébe esik:

TAX osztópontjai: 601, 1000

Négyzetlábankénti ár osztópontjai: 0.6, 0.7

Ezután számoljátok ki a diszkrét változók kereszttábláját a Beszúrás/Kimutatás parancs segítségével majd a tanult módon teszteljétek chi-négyzet próbával, hogy a diszkrét változók függetlenek-e.

Megjegyzés 1: A fenti módszer részleges függetlenségvizsgálatnak tekinthető. Ha a diszkrétizált változók összefüggnek akkor az eredeti változók is összefüggnek. Azonban a diszkrétizált változók függetlenségéből csak erősen sejthető, hogy az eredeti folytonos változók is függetlenek. A sejtés annál erősebb minél több kategóriát hozunk létre. Azonban a kategóriaszámnak határt szab, hogy ha túl sok az üres vagy közel üres cella a diszkrétizált változók kereszttáblájában akkor a chi-négyzet teszt nem működik jól (ilyenkor a khi-négyzet statisztikára a próba alapját képező aszimptotika még nem érvényesül). Ezért hoztunk létre csak 3-3 kategóriát.

Megjegyzés 2: Ha mindent jól csináltál akkor a chi-négyzet próba el fogja utasítani a diszkrétizált változók függetlenségét, vagyis az eredmény az előbbi megjegyzést is használva azt fogja mutatni, hogy a négyzetméterenkénti ár és az adó összefüggnek. Érdeemes pontfelhő ábrán ábrázolni (lineáris regressziós egyenessel és R^2 statisztikával) a két folytonos változót. A pontfelhő ábra és a kereszttábla is azt mutatja, hogy az alacsony és a magas adó kategóriában magasabb a drága házak aránya, mint a közepes adó kategóriában. Mindezt úgy interpretálhatjuk, hogy a két változó közti összefüggés nem igazán lineáris.