Tunnustevahelised seosed

Iga üldkogumi objekti võib kirjeldada mitme erineva tunnuse kaudu. Seni tegelesime ühe tunnuse uurimisega – leidsime tema keskmise, standardhälbe, jms. Sageli on aga tähtis tunnuste vaheliste seoste uurimine.

Eelnevast teame, et uuritavat tunnust käsitletakse matemaatilises statistikas kui juhuslikku suurust. Juhuslike suuruste vahel võib valitseda kas funktsionaalne sõltuvus, statistiline sõltuvus või nad võivad olla sõltumatud.

Kaks suurust on **sõltumatud**, kui ühe suuruse muutumine ei mõjuta teise suuruse muutumist. Vastasel juhul on tegemist **sõltuvate** suurustega. Kui suurused on sõltuvad ja üks suurus on täpselt leitav teise kaudu, siis on räägitakse **funktsionaalsest sõltuvusest.** Funktsionaalset sõltuvust väljendavad näiteks Ohmi seadus, Newtoni teine seadus jne. Kui üht suurust pole võimalik teise kaudu täpselt arvutada, vaid selle asemel on ühe muutuja **tendents muutuda kindlas suunas** teise muutuja muutumisel, siis on tegemist **statistilise e. stohhastilise sõltuvusega**. Stohhastilises sõltuvuses on näiteks rööv-ja saakloomade arvukus mingis piirkonnas.

Juhuslike suuruste vahelise statistilise sõltuvuse uurimiseks kasutatakse **korrelatsioon- ja regressioonanalüüsi**. **Korrelatsioonanalüüsi** kasutatakse juhuslike suuruste vahelise seose olemasolu, tugevuse ja iseloomu mõõtmiseks. Suurusi vaadeldakse sümmeetriliselt, s.t. on ükskõik kas rääkida X ja Y vahelisest või Y ja X vahelisest korrelatsioonist. Mõlemad juhuslikud suurused on üheõiguslikud vastastikku sõltuvad muutujad. **Regressioonanalüüsi** korral jäetakse niisugune sümmeetria kõrvale ja räägitakse ühe juhusliku suuruse sõltuvusest teisest juhuslikust suurusest. Mis tähendab, et ühte kahest muutujast loetakse sõltumatuks ja see väljendab põhjust ning teine muutuja tagajärge.

Korrelatsioonanalüüs

Kahe tunnuse vahelise seose kirjeldamist alustatakse sageli tunnustevahelise hajuvusdiagrammi joonistamisega. **Hajuvusdiagramm** on joonis, millele kantakse punktidena kõik valimi elemendid. Valimi i-ndale elemendile vastava punkti esimeseks koordinaadiks on esimese tunnuse väärtus x_i, teiseks koordinaadiks aga teise tunnuse väärtus y_i. Niisugust kahe tunnuse jaotuse graafikut nimetatakse ka **korrelatsioonivälja**k. Punktide paiknemine korrelatsioonivälja annab piltliku ettekujutuse tunnuste ühisest käitumisest (iseloomustab tunnuste X ja Y omavahelist seost).

Tasub meeles pidada kolme tüüpilist hajuvusdiagrammi, millele võib tulemuste kirjeldamisel tugineda. Need on järgmised

Kasvav seos | Kahanev seos | Sõltumatud tunnused

Kasvava seose korral kaasnevad ühe tunnuse suurte väärtustega enamasti teise tunnuse suured väärtused ja ühe tunnuse väikeste väärtustega enamasti teise tunnuse väikesed väärtused. **Kahaneva seose** korral kaasnevad ühe tunnuse suurte väärtustega enamasti teise tunnuse väikesed väärtused ja ühe tunnuse väikeste väärtustega enamasti teise tunnuse suured väärtused. **Sõltumatute tunnuste** korral ei mõjuta ühe tunnuse väärtus mingil moel teise tunnuse käitumist. Punktiparv on ühtlaselt hajutatud.

Lineaarne korrelatsioonikordaja

Üldiselt on hajuvusdiagrammi põhjal antud kirjeldus tunnuste ühisele käitumisele subjektiivne. Kasulik oleks omada arvulist kordajat, mis annab hinnangu tunnuste vahelise seose tugevusele ja iseloomule. Kõige sagedamini kasutatakse kahe tunnuse vahelise seose tugevuse iseloomustamiseks **lineaarset korrelatsioonikordajat** e. **Pearsoni korrelatsioonikordajat**

$$r = \frac{n\sum x_i y_i - \sum x_i \cdot \sum y_i}{\left\{\left[n\sum x_i^2 - \left(\sum x_i\right)^2\right]\cdot\left[n\sum y_i^2 - \left(\sum y_i\right)^2\right]\right\}^{1/2}}.$$

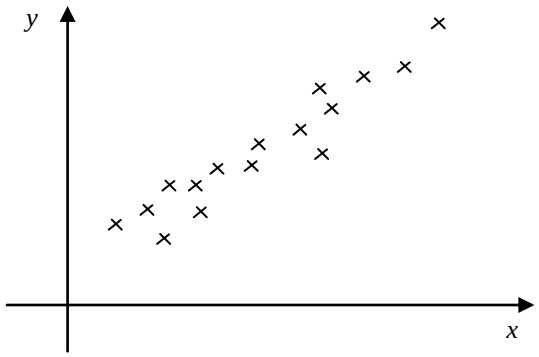
(Selle valemi abil on korrelatsioonikordaja arvutamine läbi viidud materjalidele lisatud Exceli failis.)

Lineaarsel korrelatsioonikordajal on järgmised omadused:

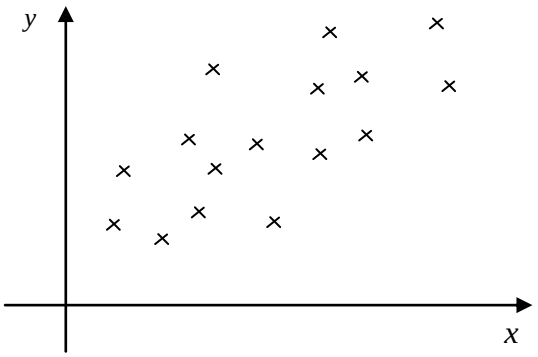
1) $-1 \le r \le 1.$
2) Kui tunnused *X* ja *Y* on sõltumatud, siis $r = 0.$ (Vastupidine ei tarvitse kehtida.)
3) Kui tunnuste vahel on lineaarne seos (s.t. kõik vaatlusandmed asuvad täpselt sirgel $Y = a + bX$, kus *a* ja *b* on konstandid), siis $|r| = 1.$

Järeldus Kui $|r|$ on lähedane ühele, siis *X* ja *Y* vaheline sõltuvus on lähedane lineaarsele.

Kui $r > 0,$ räägitakse **positiivsest korrelatsioonist**, see tähendab, et juhuslikel suurustel *X* ja *Y* on tendents muutuda samas suunas (punktid moodustavad „tõusva“ parve). Mida suurem on r , seda tugevam on tunnuste *X* ja *Y* vaheline seos. Järgnevatel joonistel on kujutatud tugevat ja nõrka positiivset korrelatsiooni.

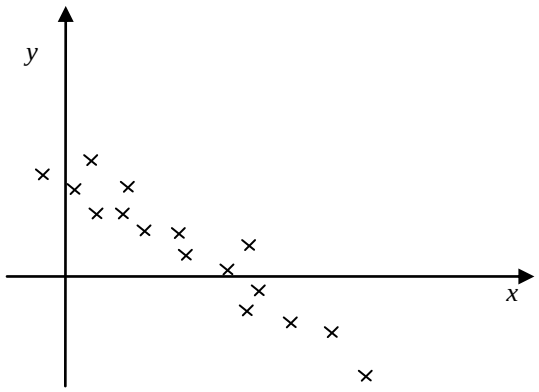


Tugev positiivne korrelatsioon



Nõrk positiivne korrelatsioon

Negatiivse korrelatsiooni korral $(r < 0)$ on ühe suuruse kasvamisel teisel suurusel tendents kahaneda.



Tugev negatiivne korrelatsioon

Oletame, et valimi abil arvutatud korrelatsioonikordaja $r \neq 0$. Kuna valim on juhuslik, siis ei saa siit veel järeldada, et ka üldkogumi korrelatsioonikordaja $r_{xy} \neq 0$. Seepärast viiakse lisaks läbi korrelatiivse sõltuvuse olemasolu kontroll hüpoteeside abil.

Püstitatakse hüpoteesid:

$H_0: r_{xy} = 0$ (vastavate üldkogumi tunnuste vahel pole lineaarset seost)

$H_1: r_{xy} \neq 0$ (vastavate üldkogumi tunnuste vahel on lineaarne seos)

1) Teststatistikuna kasutatakse sellisel juhul statistikut

$$t = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Saab näidata, et kui uuritavad tunnused on normaaljaotusega, siis nullhüpoteesi kehtides on see statistik ligikaudselt Studenti jaotusega vabadusastmete arvuga $k = n-2$.

2) Ette antakse olulisuse nivoo α ning leitakse kriitiline punkt *t*-jaotuse kvantiilide tabelist

$$t_{kr} = t\left(k, 1-\frac{\alpha}{2}\right),$$

(või *MS Exceli* keskkonnas kasutades funktsiooni *TINV*(α; *k*)).

3) Kui $|t_{emp}| < t_{kr}$ siis jäädakse nullhüpoteesi juurde (ehkki seos valimi elementide vahel on olemas, pole meil kindlust, kas see seos on üldistatav üldkogumile). Vastupidisel juhul võetakse vastu alternatiivne hüpotees (võime lugeda tõestatuks, et ka üldkogumis erineb korrelatsioonikordaja nullist).

Näide. Järgnevas tabelis on antud 8 gümnaasiumi õpilase pea ümbermõõt ja keskmine hinne:

Peaümbermõõt	54	59	60	60	60	57	65	60
Keskmine hinne	4,7	4	4,5	4,3	4	4,5	4,6	4,5

Leida lineaarne korrelatsioonikordaja ja kontrollida hüpoteesi korrelatsiooni olemasolust olulisuse nivool 0,05.

Lahendus on toodud lisana Exceli failis.

Spearmani korrelatsioonikordaja

Lineaarne korrelatsioonikordaja ei ole „kõikvõimas" seosenäitaja. Tuleb arvestada, et ta on lineaarse seose näitaja ja ta võib anda seose puudumise, kuigi tunnuste vahel on mingi teise funktsiooniga avaldatav seos.

Arv- või järjestatud tunnuste X ja Y vahelise seose monotoonsuse astme mõõtmiseks on defineeritud mitmesuguseid astak-korrelatsioonikordajaid. (Sõltuvust nimetatakse monotoonseks, kui ühe tunnuse kasvamine toob kaasa teise tunnuse kasvamise ja ühe tunnuse kahanemine toob kaasa teise tunnuse kahanemise). Astakkorrelatsioonikordaja kasutab tunnuste väärtuste asemel nende astakuid (järjekorranumbreid). Üks tuntuim nendest on Spearmani astakkorrelatsioonikordaja.

Spearmani korrelatsioonikordaja arvutamiseks tuleb:

1) Korrastada ühe tunnuse väärtused x_i järjestatud hulka, alustades vähimast ja lõpetades suurimaga.
2) Nummerdada saadud järjestatud hulga elemendid, alustades ühest (öeldakse elemendile x_i omistatakse astak r_i). Kui X väärtuste seas on korduvaid, siis võetakse neile vastavateks astakuteks esialgsete astakute aritmeetiline keskmine.
3) Omistada teise tunnuse Y väärtustele astakud q_i.
4) Arvutada summa $\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (r_i - q_i)^2$.
5) Leida Spearmani korrelatsioonikordaja $r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$.

Spearmani korrelatsioonikordaja omadusi

1) $-1 \le r_s \le 1$.
2) Kui $r_s = 1$, on tegemist range positiivse korrelatsiooniga (kui kasvab X, siis kasvab alati ka Y, kuid see kasv ei tarvitse olla lineaarne).
3) Kui $r_s = -1$, on tegemist range negatiivse korrelatsiooniga (ühe tunnuse kasvamisele vastab teise tunnuse kahanemine).
4) Kui $r_s = 0$, siis on tunnused Spearmani mõttes mittekorreleeruvad.

Korrelatiivse sõltuvuse olemasolu kontrollitakse täpselt sama algoritmi alusel, kui lineaarse korrelatsioonikordaja korral.

Näide. Järgnevas tabelis on antud andmed hiirte kohta.

Hiirte pikkus	11,1	12	10,6	9,8	10,2	10,5	11,6	10,6	10	10,2	11,5	10,5	15,9
Hiirte kaal	22	25	23	20	20	22	22	25	25	23	28	21	36

Leida Spearmani korrelatsioonikordaja ja kontrollida hüpoteesi korrelatsiooni olemasolust olulisuse nivool 0,05.

Lahendus on toodud lisana Exceli failis.