Korrelatiivne sõltuvus

Põhjuslik sõltuvus, statistiline sõltuvus, sõltumatus, hajuvusdiagramm, lineaarne mudel, regressioonisirge*, korrelatsioonikordaja, korreleeritud tunnused, mittekorreleeritud tunnused, sõltuvuse suund,

Gümnaasium

Sõltuvuse liigid. Põhjuslik sõltuvus

Looduses ja ühiskonnas toimuvad protsessid ja nähtused mõjutavad üksteist vastastikku, sel korral räägitakse sõltuvatest sündmustest, sõltuvatest juhuslikest suurustest ja sõltuvatest protsessidest. Sõltuvusi on mitmesuguseid. Kõige tuntum on põhjuslik sõltuvus (seos).

Kahe **sündmuse** vahel on põhjuslik sõltuvus, kui üks sündmus on teise tagajärg, st et ühe sündmuse toimumisest järeldub kindlasti teise sündmuse toimumine.

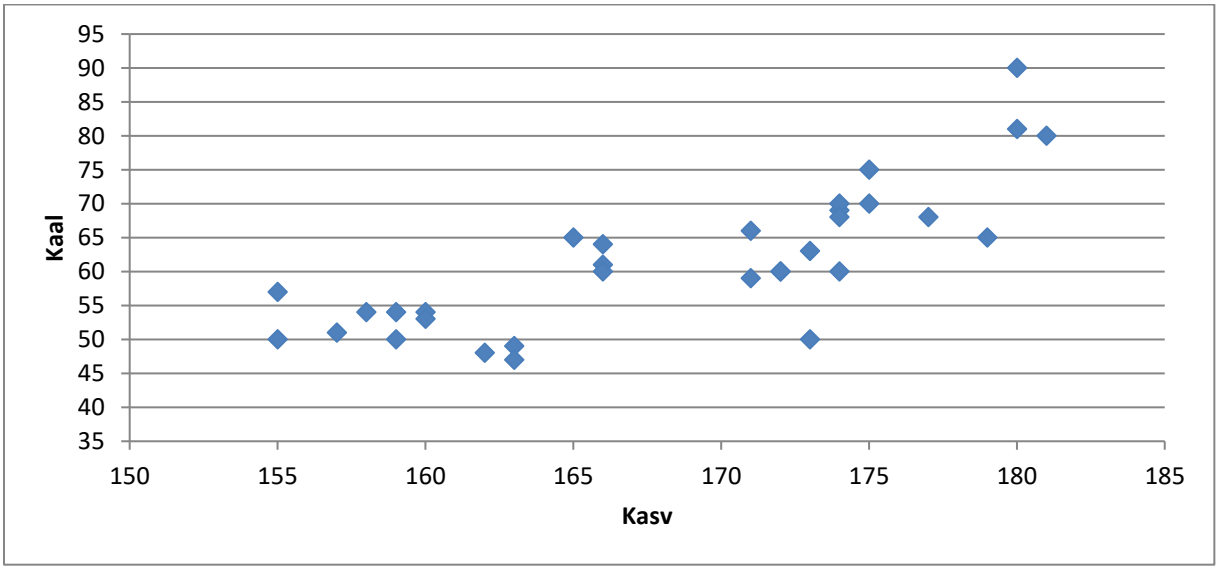
Statistiline sõltuvus ja ennustamine

Kui ühe sündmuse toimumise tulemusena teine sündmus võib toimuda, kuid võib ka mitte toimuda, siis nende sündmuste vahel põhjuslikku seost ei ole, kuid võib olla **statistiline sõltuvus.** Sel juhul on **sündmused sõltuvad**. Statistiline sõltuvus annab võimaluse ühe sündmuse toimumise põhjal **ennustada** teise sündmuse toimumist või mittetoimumist. Kui sündmuste vahel statistilist sõltuvust ei ole, siis on need **sündmused sõltumatud** (vt Sündmuste sõltuvus).

Ka juhuslike suuruste (tunnuste) vahel võib olla nii põhjuslik kui ka statistiline sõltuvus, kuid tunnused võivad olla ka sõltumatud. **Statistiline sõltuvus** on statistika üks põhimõisteid. Statistiline sõltuvus on eeldus selleks, et uuritavaid andmeid saaks kasutada nähtuste seletamiseks ja ennustamiseks. Kui kahe juhusliku suuruse vahel on **statistiline sõltuvus**, siis võimaldab ühe juhusliku suuruse väärtuse teadmine ennustada (rohkem või vähem täpselt) teise tunnuse väärtusi. Juhuslike suuruste vahelise statistilise sõltuvuse tuntuim, kuid mitte ainus liik on **korrelatsioon (korrelatiivne sõltuvus).**

Hajuvusdiagramm

Näitena vaatame klassi õpilaste kasvu ja kaalu sõltuvust, mida illustreerib **hajuvusdiagramm**



Jooniselt on näha, et õpilaste pikkus ja kaal ei muutu juhuslikult, vaid pikemad kaaluvad üldiselt rohkem ja lühemad vähem. Kasvu ja kaalu vahel on **statistiline sõltuvus**. Statistiline sõltuvus võib väljendada nähtuste põhjuslikku seost, mis võib-olla kas otsene (üks tunnus mõjutab teist), kuid võib ka tuleneda mingi kolmanda tunnuse mõjust mõlemale tunnusele.

Käesolevas näites kindlasti mõjutab pikkus kaalu, kuid niihästi pikkus kui kaal sõltuvad isiku omapärast, muuhulgas ka sellest, kas mõõdetav isik on poiss või tütarlaps.

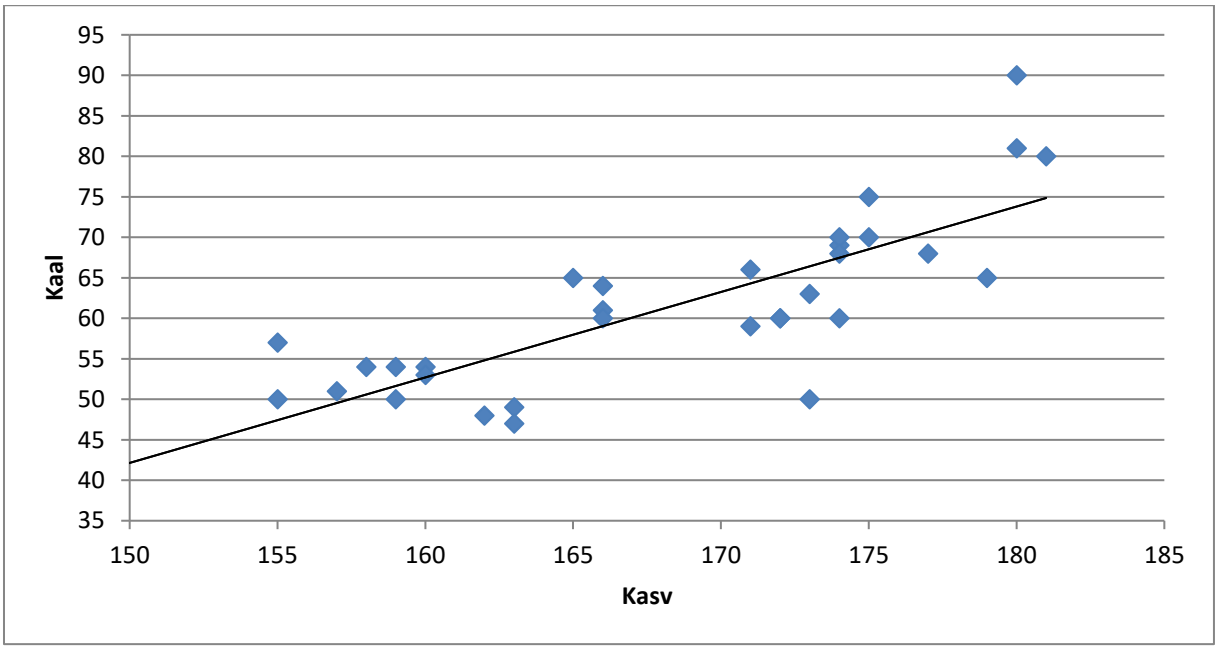
Uuritavaid tunnuseid siduv mudel

Kui kahe tunnuse vahel on statistiline sõltuvus, siis on võimalik konstrueerida mudel ühe tunnuse väärtuse prognoosimiseks teise järgi. See mudel erineb seni käsitletud mudelitest, mis määrasid mingi tunnuse jaoks sobiva tõenäosusjaotuse. Praegu loodav mudel seob kaht tunnust omavahel, kuid ka selle mudeli jaoks on tarvis valimi põhjal määrata parameetrid.

Kõige lihtsam mudel on lineaarne, st et mudelit kirjeldab sirgjoon.

Lineaarne mudel

Alljärgnevale joonisele on kantud sirge, mis esindab kaalu lineaarset prognoosi kasvu järgi (antud klassi õpilaste andmetel). Seda sirget nimetatakse **regressioonijooneks**. Seda võib tõlgendada nii, et valides horisontaalteljelt pikkuse, saame sellele pikkusele vastava keskmise kaalu regressioonijoonelt. Näiteks 160 cm pikkuse õpilase puhul oleks keskmine kaal (antud mudeli kohaselt) 53 kg, 170 cm pikkuse õpilase kaal peaks mudeli kohaselt olema 63 kg. Punktid, mis paiknevad ülevalpool regressioonijoont, vastavad õpilastele, kelle kaal on suurem kui keskmine nende pikkusele vastav kaal selle klassi andmetel. Regressioonijoonest allpool asuvad punktid aga näitavad õpilasi, kelle kaal on mudeliga määratud kaalust väiksem.



Regressioonijoone võrrand (mudel) leitakse valimi punktide x- ja y-koordinaatide järgi, kusjuures see on sirge, mis kõige paremini kirjeldab y (kaalu) sõltuvust x-st (kasvust). Sirgjoone jaoks on tarvis määrata kaks parameetrit: tõus ja algordinaat (vabaliige).

Käesoleval juhul on mudeli võrrand alljärgnev:

Y = -116,13 + 1,055x.

Siit on näha, et keskmiselt lisab iga pikkuse sentimeeter pisut üle ühe kilogrammi kaalu. Kuna pikkus sentimeetrites on enam kui 100 ühiku võrra suurem kui kaal kilodes, on vabaliige selle vahekorra tasakaalustamiseks negatiivne.

Mudelit on mõtet kasutada ühe tunnuse väärtuste prognoosimiseks teise tunnuse väärtuste järgi, vaid siis, kui tunnuste vahel on küllalt tugev sõltuvus. Tunnustevahelise korrelatiivse Sõltuvuse tugevust ehk lineaarse mudeli headust mõõdab **korrelatsioonikordaja.**

Korrelatsioonikordaja

Korrelatsioonikordajal on alljärgnevad olulised omadused:

- Korrelatsioonikordaja väärtus võib muutuda –1 ja 1 vahel.
- Korrelatsioonikordaja on positiivne siis, kui ühe tunnuse suurenedes ka teine tunnus suureneb (nii, nagu on olukord kasvu ja kaalu puhul). Sel juhul regressioonisirge on tõusev.
- Korrelatsioonikordaja on negatiivne siis, kui ühe tunnuse kasvades teine tunnus kahaneb. Näiteks, kui üks tunnus on spordile kulutatud aeg ja teine tunnus – teleri vaatamisele kulutatud aeg, siis neid iseloomustava mudeli puhul on korrelatsioonikordaja negatiivne ja regressioonisirge on langev.
- Kui tunnustevaheline seos on tugev, siis paiknevad kõik punktid regressioonijoonte lähedal ja korrelatsioonikordaja on lähedane arvule 1 või –1 vastavalt sõltuvuse suunale. Sel juhul saab ühe tunnuse väärtusi üsna täpselt teise tunnuse väärtuste järgi ennustada.
- Kui tunnustevaheline sõltuvus on nõrk, siis paiknevad punktid hajuvusdiagrammil hajusalt üle kogu diagrammi ja korrelatsioonikordaja on nulli lähedane. Sel juhul ei anna ühe tunnuse väärtuste teadmine olulist teavet teise tunnuse väärtuste kohta, prognoosimine on tulutu.
- Korrelatiivne sõltuvus on vastastikune – juhuslikku suurust X saab Y järgi sama täpselt prognoosida nagu juhuslikku suurust Y X järgi.
- Korrelatiivsest sõltuvusest kahe juhusliku suuruse vahel ei saa järeldada, et nende suuruste vahel oleks põhjuslik seos.
- Kui tunnuste vaheline põhjuslik seos on olemas, on nende vahel alati ka korrelatiivne sõltuvus, kuid korrelatsioonikordaja ei näita seda, missugune on selle suund, st et kas X mõjutab Y või vastupidi.
- Korrelatsioonikordaja ruut näitab, missuguse osa ühe tunnuse hajuvusest kirjeldab lineaarne mudel. Käesolevas näites on korrelatsioonikordaja väärtus 0,82, see on väga tugev korrelatsioon. Korrelatsioonikordaja ruut on 0,67, seega kirjeldab kasv lineaarse mudeli abil kaalu hajuvust 67% ulatuses.

Valemid

- Olgu valimi maht n ja selle punktid $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- Valimi keskmised olgu $\bar{x} = (x_1+x_2+\ldots x_n)/n$ ja $\bar{y} = (y_1+y_2+ \ldots +y_n)/n$
- Valimi standardhälbed olgu vastavalt s(x) ja s(y),
- $s(x) = \sqrt{[(x_1- \bar{x})^2 + (x_2- \bar{x})^2 + \ldots + (x_n- \bar{x})^2]/(n—1)}$,
- $s(y) = \sqrt{[(y_1- \bar{y})^2 + (y_2-\bar{y})^2 + \ldots + (y_n- \bar{y})^2]/(n—1)}$.

Siis **lineaarse mudeli Y = aX + b parameetrid** arvutatakse valemitest:

$$a = [(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \ldots + (x_n - \bar{x})(y_n - \bar{y})] / [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2],$$

$$b = \bar{y} - a\,\bar{x}.$$

Korrelatsioonikordaja r

arvutatakse valemist

$$r = [(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \ldots + (x_n - \bar{x})(y_n - \bar{y})] / [(n—1)s(x)\, s(y)).$$

Kõnekeeles kasutatakse korrelatsiooni vahel ka statistilise sõltuvuse sünonüümina. Enamasti ei põhjusta see eksitusi, sest alati, kui juhuslikud suurused on korreleeritud, on nende vahel statistiline sõltuvus. Aga sellest, et juhuslikud suurused on mittekorreleeritud, ei järeldu nende sõltumatus.