Matemaatiline statistika

Matemaatiline statistika on matemaatika haru, mis käsitleb statistiliste andmete põhjal järelduste tegemise meetodeid. Statistiliste andmete hulka kuuluvad näiteks vaatlus-, mõõtmis-, katse- ja küsitlustulemused. Matemaatilise statistika aluseks on tõenäosusteooria.

Statistika põhimõisted

Tutvume järgnevalt mõnede statistika põhimõistetega.

Üldkogum on vaatluse alla võetav objektide koguhulk, näiteks kogu Eesti elanikkond, kõikide Eesti üliõpilaste hulk, kõikide raamatute hulk raamatukogus jne.

Uurimisobjektide vaadeldavaid omadusi nimetatakse statistikas **tunnusteks.** Uuritavat tunnust käsitletakse matemaatilises statistikas kui juhuslikku suurust, mis võib uurijale tundmatute ja kontrollimatute faktorite mõjul omandada erinevaid väärtusi. Näiteks tudengi tunnusteks võivad olla tema keskmine hinne, sissetulek kuus, telefoniarve kuus, pikkus, kaal, rahvus, silmade värv jne. Sõltuvalt tunnuste väärtuste iseloomust jagunevad tunnused kvantitatiivseteks ja kvalitatiivseteks. **Kvantitatiivsed tunnused** on arvulised. Näiteks tudengi sissetulek kuus, pikkus, kaal jne. **Kvalitatiivsel tunnusel** ei ole arvulisi näitajaid. Kvalitatiivsed tunnused on näiteks sugu, rahvus, silmade värv.

Tunnuste varieeruvuse mõõtmiseks kasutatakse nelja skaalat: nominaalskaala, ordinaalskaala, intervallskaala ja meetriline skaala.

Nominaalskaala koosneb üksteisest sõltumatutest klassidest, mida ei saa loogiliselt järjestada, kusjuures iga objekt saab kuuluda ainult ühte klassi (rahvus, sugu, silmade värv). Nominaalskaala kasutamise korral nimetatakse mõõtmist tavaliselt klassifitseerimiseks.

Ordinaalskaala koosneb erinevatest klassidest, mida on võimalik järjestada. Näiteks eksamitulemus “puudulik”, “kasin”, ..., “suurepärane”; küsitlustulemused “vastu”, “pigem vastu kui poolt”, ..., “poolt”.

Nominaal- ja ordinaalskaalat kasutatakse kvalitatiivsete tunnuste mõõtmiseks.

Intervallskaala määrab üksikute mõõtmistulemuste vahelise kauguse, kusjuures fikseeritud nullpunkti ei eksisteeri ning suhtarve ei saa moodustada. Näiteks temperatuur Celsiuse või Fahrenheiti kraadides (ei saa öelda, et 20°C on kaks korda soojem kui 10°C).

Meetriline skaala määrab üksikute mõõtmistulemuste vahelise kauguse, kusjuures eksisteerib fikseeritud nullpunkt. Meetriliste andmete jaoks saab moodustada suhtarve. Näiteks võib öelda, et kahetunnine ajavahemik on neli korda pikem pooletunnisest.

Meetriliselt skaalalt võib vajaduse korral üle minna intervallskaalale, rühmitades võimalikud väärtused klassidesse. Näiteks kui kogume andmed rahvastiku kohta, fikseerides iga inimese kohta vanuse, siis oleme saanud tulemused meetrilisel skaalal. Kui rühmitame kogutud andmed 20 kaupa (0-20; 20-40; …), siis see tähendab, et oleme teisendanud tulemused intervallskaalale.

Üldkogumi uurimisel on kaks võimalust: uurida üldkogumi kõiki elemente või ainult teatud osahulka. Praktikas kasutatakse kõikset uurimist väga harva. Tavaliselt piirdutakse üldkogumi mingi juhusliku alamhulga - **valimi** vaatlusega. Valimi põhjal tehakse järeldusi üldkogumi kohta. Valimi moodustamisel peavad olema täidetud kaks järgmist nõuet:

1. valimi maht peab olema küllalt suur;
2. igal üldkogumi indiviidil peab olema võrdne võimalus sattuda valimisse.

Neid kaht nõuet rahuldavat valimit nimetatakse **representatiivseks** e. **esindavaks**. Nende nõuete vajalikkus on ilmne: liiga väikeste valimite puhul ei saa me teha usaldusväärseid otsustusi üldkogumi kohta; kui aga valimi moodustamisel eelistatakse teatud indiviide teistele (soovides

leida Eestis elavate meeste keskmist pikkust viime mõõtmised läbi vaid korvpallurite hulgas), siis see valim võib anda moonutatud pildi üldkogumist.
Vaatluse tulemusel saadakse harilikult **korrastamata statistiline rida**, milles andmed paiknevad registreerimise või laekumise järjekorras. Andmete paigutamisel kasvavasse või kahanevasse järjekorda, saadakse **korrastatud rida**, mida matemaatilises statistikas nimetatakse ka **variatsioonreaks**.
Juhul kui valimis mahuga n on võrdseid elemente (väärtus x_i, esineb n_i korda), siis esitatakse variatsioonrida kujul

x_i	x_1	x_2	...	x_m
$p_i^* = \frac{n_i}{n}$	$\frac{n_1}{n}$	$\frac{n_2}{n}$	...	$\frac{n_m}{n}$

Variatsioonrea lühendamiseks rühmitatakse elemendid sageli klassidesse

x_i	$[a_0\,;\,a_1)$	$[a_1\,;\,a_2)$	...	$[a_{m-1}\,;\,a_m]$
$p_i^* = \frac{n_i}{n}$	$\frac{n_1}{n}$	$\frac{n_2}{n}$	...	$\frac{n_m}{n}$

Võimaluse korral valitakse kõik klassid ühesuguse ulatusega. Enne andmete rühmitamist tuleb valida sobiv klasside arv ja esimese klassi algväärtus.
Liiga väikese klasside arvu korral kaotatakse liiga palju informatsiooni juhusliku suuruse kohta. Liiga suure klasside arvu korral võib tekkida raskusi tulemuste üldistamisega. Üldiselt soovitatakse valida klassi intervall nii, et valimi, mille maht on n, rühmitamise tulemusel saadakse $m = 1 + \lceil \log_2 n \rceil$ klassi[1].
Olles fikseerinud valimi ning moodustanud mingit tunnust mõõtes variatsioonrea, saame moodustada **üldkogumi empiirilise jaotusfunktsiooni**:

$$F^*(x) = P(X^* < x) = \sum_{x_i < x} \frac{n_i}{n},$$

kus X^* on diskreetne juhuslik suurus, mille jaotustabel on moodustatud variatsioonrea abil.

On tõestatud, et valimi mahu n tõkestamatu kasvamise korral koondub empiiriline jaotusfunktsioon $F^*(x)$ tõenäosuse järgi üldkogumi jaotusfunktsiooniks $F(x)$.

Empiirilise jaotusfunktsiooni graafikuks on treppjoon, mis meenutab diskreetse juhusliku suuruse jaotusfunktsiooni graafikut.

Näide. On antud 50-aastase männiku 60 juhuslikult valitud puu kõrgused, mis on mõõdetud 0,1 m täpsusega. Rühmitada andmed, joonestada empiirilise jaotusfunktsiooni graafik ja histogramm. Ülesande andmed ja lahendus on toodud lisana esitatud Exceli failis.

[1] $\lceil a \rceil$ - tähistab arvu a täisosa

Statistilised hinnangud

Kui valim on moodustatud, siis selle parameetrite (keskväärtus, dispersioon jm.) arvutamine ei valmista raskusi. Et me valimi kaudu soovime iseloomustada üldkogumit, siis huvitavad meid tegelikult üldkogumi kohta käivad arvulised karakteristikud. Neid saaksime otseselt arvutada ainult siis, kui kõik üldkogumi elemendid oleksid teada; selline olukord esineb statistikas vaid erandjuhtudel. Seepärast püstitamegi küsimuse, kui hästi iseloomustavad valimi arvulised karakteristikud üldkogumit. Esitatud küsimuses peitub statistiliste hinnangute probleem. Kõik statistilised hinnangud jaotatakse punkt- ja vahemikhinnanguteks. Järgnevalt tutvume punkthinnangutega.

Punkthinnangud

Olgu antud juhuslik suurus *X*, mille jaotust iseloomustab parameeter *a* (väärtus on tundmata). Võtame mingi valimi, mille korral see juhuslik suurus omandab väärtused $x_1, x_2,..., x_n$ ja arvutame selle jaoks parameetri väärtuse $\tilde{a}$. Igale valimile vastab üldiselt erinev $\tilde{a}$, seega võime kirjutada

$$\tilde{a} = \tilde{a}(x_1, x_2,..., x_n).$$

Väärtus $\tilde{a}$ ongi parameetri *a* **punkthinnanguks**.

Et see hinnang iseloomustaks võimalikult hästi üldkogumi vastavat parameetrit, peavad olema täidetud 3 nõuet:

1. **hinnangu nihutamatus**

Valimi põhjal arvutatud parameetrit nimetatakse üldkogumi karakteristiku nihutamata hinnanguks, kui tema keskväärtus on võrdne hinnatava parameetriga, s.t.

$$E\left[\tilde{a}(x_1, x_2,..., x_n)\right] = a.$$

Kui viimane võrdus pole täidetud siis nimetatakse hinnangut **nihutatuks** e. **nihkega hinnanguks**. Nihutatud hinnanguid pole soovitatav kasutada, kuna need tingivad süstemaatilise vea.

2. **hinnangu efektiivsus**

Hinnangut nimetatakse **efektiivseks**, kui ta on nihutamata ja tema dispersioon on minimaalne. Efektiivsed hinnangud võimaldavad saavutada vajalikku täpsust kõige väiksema mahuga valimite korral.

3. **hinnangu konsistentsus (mõjusus, sisukus)**

Hinnangut $\tilde{a}$ nimetatakse **konsistentseks**, kui ta koondub tõenäosuse järgi parameetriks *a*.

Saab tõestada, et üldkogumi keskväärtuse efektiivseks nihutamata ja konsistentseks hinnanguks on valimi **aritmeetiline keskmine** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Samuti võib veenduda, et üldkogumi dispersiooni hinnanguks sobib suurus

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

ja standardhälbe hinnanguks

$$s = \frac{1}{\sqrt{n-1}}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Klassidesse jagatud valimi korral tuleb aritmeetilise keskmise ja valimdispersiooni leidmiseks kasutada klasse määravate vahemike keskpunkte. Olgu variatsioonirida esitatud kujul

x_i	$[a_0\,;\,a_1)$	$[a_1\,;\,a_2)$	...	$[a_{m-1}\,;\,a_m]$
$p_i^* = \frac{n_i}{n}$	$\frac{n_1}{n}$	$\frac{n_2}{n}$	...	$\frac{n_m}{n}$

Siis

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{m} n_i \cdot x_i^k\,,$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{m} n_i \cdot (x_i^k - \bar{x})^2\,,$$

kus $x_i^k = \frac{a_{i-1} + a_i}{2}$.

Näide. Mõõdeti 270 detaili. Nende pikkus on vahemikus 66 cm kuni 90 cm. Antud on klassifitseerimisel saadud variatsioonrida. Leida jaotuse keskväärtuse, dispersiooni ja standardhälbe nihutamata hinnangud.
Ülesande andmed ja lahendus on toodud lisana esitatud Exceli failis.