Peatükk 1

Sissejuhatus

1.1 Mitteparameetrilise statistika mõistest

Mitteparameetrilise statistika mõiste on lai ja kaks mitteparameetrilisele statistikale pühendatud raamatut võivad käsitleda (esmapilgul) täiesti erinevaid valdkondi. Sestap alustuseks paari sõnaga sellest, mida mõeldakse mitteparameetrilise statistika ja mitteparameetriliste meetodite all.

Klassikalises statistikas nõutakse sageli, et me teaksime uuritava tunnuse kohta eelinformatsiooni — näiteks seda, millisesse jaotuste perre ta kuulub (kas tegemist on normaaljaotusega, eksponentjaotusega, gammajaotusega või hoopis ...-jaotusega). Teades jaotuste peret (näiteks normaaljaotus) piisab vaid paari parameetri (keskväärtuse, dispersiooni) hindamisest uuritava tunnuse jaotuse üheseks määramiseks. Teades neid paari parameetrit, saame vastata mistahes küsimusele uuritava tunnuse jaotuse kohta. Sestap pöörabki klassikaline parameetriline statistika suurt tähelepanu jaotusparameetrite hindamisele ja nende kohta hüpoteeside kontrollimisele.

Paraku praktikas osutub kitsa parameetrilise jaotuste pere täpne määramine sageli vägagi raskeks. Seega praktiliste ülesannete lahendamisel puututakse sageli kokku nn mitteparameetriliste jaotuste peredega.

Definitsioon 1.1 *Jaotuste peret kutsutakse* ***mitteparameetriliseks jaotuste pereks****, kui vastavasse perre kuuluvat jaotust ei saa üheselt identifitseerida kasutades vaid vähest arvu parameetreid (üheks selliseks jaotuste pereks on näiteks kõigi pidevate jaotuste pere).*

Statistilist meetodit, mis on rakendatav tunnus(t)e uurimiseks, mille jaotus kuulub mitteparameetriliste jaotuste perre, nimetatakse **mitteparameetriliseks meetodiks**. Juhul, kui tahetakse rõhutada, et üks või teine meetod

ei eelda mõnda lihtsat jaotust (nagu normaaljaotust või beetajaotust) siis räägitakse ka jaotusvabadest (*distribution-free*) meetoditest.

Vaatame ühte teist näidet. Tunneme huvi, milline on tunnuse Y tinglik jaotus juhul, kui teame tunnust X, $f_{Y|X}$. Oletame, et Y-tunnuse tinglikuks jaotuseks on normaaljaotus. Tavapärane, parameetriline lähenemine nõuab, et teaksime tunnuste Y ja X vahelist seost piisavalt täpselt, et kõigest peale mõne täiendava parameetri hindamist saame teada seose tunnuste vahel. Näiteks eeldame, et $Y|X \sim N(c_0 + c_1 X, \sigma^2)$. Niipea kui saame teada parameetrite c_0 ja c_1 väärtused (või hindame need), teame ka seost tunnuste vahel.

Vahel pole aga seost tunnuste vahel võimalik piisavalt täpselt eelnevalt kirja panna. Mõnikord saame kõigest öelda, et $Y|X \sim N(f(X), \sigma^2)$, kus $f()$ on näiteks mingi suvaline pidev funktsioon. Sellisel juhul pole võimalik määratleda uuritava tunnuse (tinglikku) jaotust vaid paari tundmatu parameetri hindamise abil. Sestap kutsutakse ka toodud näites kahe tunnuse vahelise seose leidmist mitteparameetriliseks regressiooniks. Kui soovitakse rõhutada, et mõned eeldused siiski on tehtud (normaaljaotus), aga midagi pole suudetud/osatud/tahetud parametriseerida (funktsiooni f()), siis räägitakse ka semiparameetrilisest (*semiparametric*) või poolparameetrilisest meetodist.

Järgnevalt vaatame mõningaid näiteid, kuidas erinevatele küsimustele saab leida vastuseid mitteparameetrilisi meetodeid kasutades.

1.2 Tihedusfunktsiooni hindamine

Tihedusfunktsiooni hindamisest räägime pikemalt hiljem. Siin toome lihtsalt ühe näite, rõhutamaks parameetrilise ja mitteparameetrilise statistika erinevust. Parameetrilise statistika puhul eeldatakse, et vaatlused on mingist jaotuste perest (näiteks eeldatakse, et tegemist on normaaljaotusega), hinnatakse parameetrid — keskväärtust hinnatakse näiteks valimi keskmisega ja populatsiooni dispersiooni saab hinnata valimi dispersiooniga ning ongi meil leitud hinnang uuritava tunnuse tihedusfunktsioonile. Mitteparameetrilise meetodi näitena võib aga tuua histogrammi. Vaata näiteks joonist 1.1.

1.3 Märgitest

Paljud testid (t-test ja regressioon- ning dispersioonanalüüsis kasutatav F-test, tõepärasuhtel baseeruvad testid jne) eeldavad, et me teaksime uuritava

tunnuse jaotust. Vaatame, kuidas on võimalik hüpoteese kontrollida eeldamata uuritava tunnuse jaotuse kohta midagi peale pidevuse.

Märgitest võimaldab kontrollida hüpoteese uuritava tunnuse mediaani kohta,

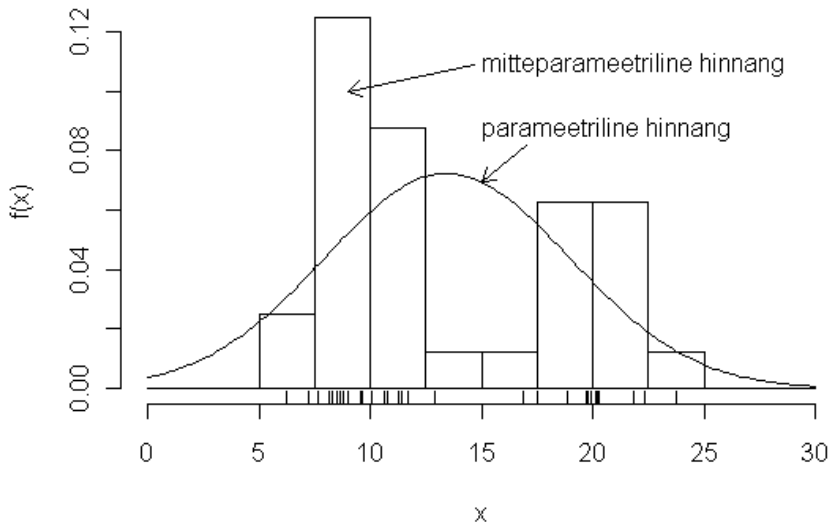
$$\begin{aligned} H_0 &: \text{med}(X) = \Theta_0 \\ H_1 &: \text{med}(X) \neq \Theta_0. \end{aligned}$$

Kuidas konstrueerida sobivat teststatistikut? Mediaan on väärtus, millest väiksemaid ja suuremaid väärtuseid peaks olema ühepalju. Seega, kui Θ_0 on tegelikult mediaan, peaks juhuslikest suurustest $X_1 - \Theta_0, \ldots, X_n - \Theta_0$ ligikaudu pooled olema suuremad nullist (vaatlus oli suurem mediaanist) ja ligikaudu pooled väiksemad nullist (vaatlus oli väiksem mediaanist). Täpsemalt öeldes:

$$P(X_i - \Theta_0 > 0) = P(X_i - \Theta_0 < 0) = 1/2,$$

kui Θ_0 on ka tegelikult mediaan. Meie teststatistik võiks siis loendada, kui mitmel korral nägime valimis Θ_0-st suuremat väärtust, $B(X, \Theta_0) = \sum_{i=1}^{n} I_{X_i - \Theta_0 > 0}$.

Joonis 1.1: Parameetriline ja mitteparameetriline hinnang tihedusfunktsioonile



Juhul kui Θ_0 on tegelik mediaan (nullhüpoteesi kehtides) siis peaks teststatistiku jaotus olema binoomjaotus, $B(X,\Theta_0) \overset{H_0}{\sim} B(n,0.5)$. Alternatiivina võime kasutada teststatistikuna suurust

$$S(X,\Theta_0) = \sum_{i=1}^{n} \mathrm{sgn}(X_i - \Theta_0).$$

Ka viimase jaotuse saame leida binoomjaotust kasutades, kui paneme tähele, et $S(X,\Theta_0) = 2B(X,\Theta_0) - n$.

Näide 1.1 *Vaatame valimit suurusega $n = 10$. Statistiku $S(X,\Theta_0)$ jaotus nullhüpoteesi $med(X) = \Theta_0$ kehtides on leitav kasutades binoomjaotust $(S(X,\Theta_0)+n)/2 \sim B(n,0.5)$ ehk $P(S(X,\Theta_0) = x) = \binom{x/2+5}{10} 0.5^{5+x/2} 0.5^{5-x/2}$. Vaata statistiku jaotust tabelist 1.1.*

Tabel 1.1: Märgitesti statistiku jaotus nullhüpoteesi kehtides, $n = 10$

x	$P(S(X,\Theta_0) = x)$
-10	0.001
-8	0.010
-6	0.044
-4	0.117
-2	0.205
0	0.246
2	0.205
4	0.117
6	0.044
8	0.010
10	0.001

Näeme, et tõenäosusega 0,89 peaks statistiku väärtused jääma vahemikku [-4..4], tõenäosusega 0,978 aga vahemikku [-6..6]. Kui statistiku $S(X,\Theta_0)$ väärtus peaks olema -6 väiksem või suurem 6-st, saame olulisuse nivool 0,05 väita, et nullhüpotees ei pea paika.

Näide 1.2 *Soovime kontrollida järgmiseid hüpoteese:*

$$\begin{aligned} H_0 &: med(X) = 4 \\ H_1 &: med(X) \neq 4. \end{aligned}$$

Olgu vaatlusandmed (järjestatud) järgmised: 3,38; 3,95; 4,15; 4,46; 4,62; 5,19; 5,44; 5,81; 5,82; 6,56. Neist kaks on väiksemad kui 4, ülejäänud on suuremad, seega

$$S(X,4) = -1-1+1+1+1+1+1+1+1+1 = 6.$$

Testi olulisustõenäosuse leidmiseks arvutame, kui tõenäoliselt võiks nullhüpoteesi kehtides näha sedavõrd ekstreemset (või veel ekstremaalsemat) teststatistiku väärtust (kasutame näites 1.1 toodud tabelit 1.1):

$$\begin{aligned} p-value &= 0,001(S=10)+0,010(S=8)+0,044(S=6)+ \\ &\quad 0,044(S=-6)+0,010(S=-8)+0,001(S=-10) \\ &= 0,11 \end{aligned}$$

Seega olulisuse nivool 0,05 peaksime jääma nullhüpoteesi juurde (mediaan võib küll olla 4).

Paneme tähele, et üsna analoogsel viisil saab kontrollida hüpoteese mistahes teise populatsiooni kvantiili kohta. Näiteks soovides testida, kas alumine kvartiil (0,25-kvantiil) võiks olla 4, peame vaid olulisustõenäosuse arvutamisel kasutama binoomjaotust $B(n, 0.25)$ (sest 0,25-kvantiilist väiksema väärtuse saamise tõenäosus on 0,25).

1.4 Usaldusintervall mediaanile

Järgneval vaatame, kuidas saab leida mitteparameetrilist usaldusintervalli mediaanile. Usaldusintervalli konstrueerimiseks saab kasutada märgitesti — $1-\alpha$-usaldusvahemikku moodustavad kõik need Θ_0 väärtused, mille puhul märgitest olulisuse nivool α otsustab nullhüpoteesi kasuks (kui $\Theta_0 = \Theta$ ehk kui nullhüpotees kehtib, siis ei tohi testi olulisustõenäosus olla väiksem α-st suurema tõenäosusega kui α, seega jääb tegelik mediaan Θ usaldusintervalli vähemalt tõenäosusega $1-\alpha$).

Näide 1.3 *Vaatame näites 1.2 toodud vaatluseid. Tabelis 1.1 on toodud statistiku $S(X,\Theta_0)$ jaotus (eeldusel, et nullhüpotees kehtib). Näeme, et olulisuse nivool 0.022 jääksime nullhüpoteesi juurde siis, kui $-6 \le S(X,\Theta_0) \le 6$. Leiame, milliste Θ_0 väärtuste korral me saame märgitesti statistiku väärtuseid vahemikust [-6..6]. See pole eriti keeruline ülesanne, sest märgitesti statistiku väärtus võib muutuda vaid valimi vaatluste poolt määratud punktides, vt ka tabelit 1.2. Selliste Θ_0 väärtuste vahemik, [3,95..5,82] olekski*

0,978-usaldusintervall mediaanile. Kuna $P(-4 \le S(X,\Theta_0) \le 4) = 0.89$*, siis 0,89-usaldusintervalliks on vahemik [4,15..5,81]. Märkame, et täpselt 0,9- või 0,95-usaldusintervalli leida pole võimalik.*

Tabel 1.2: Usaldusintervall mediaanile, märgitesti abil

Θ_0	$S(X,\Theta_0)$
$-\infty$..3,38	10
3,38..3,95	8
3,95..4,15	6
4,15..4,46	4
4,46..4,62	2
4,62..5,19	0
5,19..5,44	-2
5,44..5,81	-4
5,81..5,82	-6
5,82..6,56	-8
6,56..∞	-10

1.5 Mediaani mitteparameetriline hindamine

Mediaani oskame muidugi juba ammu valimi põhjal hinnata. Siin tutvustame lihtsalt loogikat, mida saab hiljem üldistada ka teiste situatsioonide jaoks.

Meil on statistik, antud juhul $S(X,\Theta_0)$, mille abil saame testida hüpoteese mediaani kohta. Millise statistiku $S(X,\Theta_0)$ väärtuse puhul jääme kõige kindlamini nullhüpoteesi juurde? Selliseks statistiku väärtuseks oleks $S(X,\Theta_0) = 0$, mille puhul hüpoteesi $H_0 : \Theta = \Theta_0$ olulisustõenäosuseks tuleks 1. Seega kõige paremini sobivaks hinnanguks Θ väärtusele on arv $\tilde{\Theta}$, mille puhul $S(X,\Theta) = 0$. Teine argument, millega jõutakse sama tulemuseni, kõlab järgmiselt. Kui Θ on tegelik mediaan, siis $\mathrm{E}S(X,\Theta) = 0$ (sest binoomjaotus on sümmeetriline) ja momentide meetodit kasutades võiksime valida mediaani hinnanguks $\tilde{\Theta}$ sellise väärtuse, mille puhul $S(X,\tilde{\Theta}) = 0$ ehk $\tilde{\Theta} \in (4,62 \ldots 5,19)$. Millist konkreetset väärtust antud vahemikust valida punkthinnanguks, on juba kokkuleppe küsimus. Üheks enam-vähem loogiliseks valikuvõimaluseks oleks valida lõigu keskpunkt $(4,62+5,19)/2 = 4,905$, milline olekski üks võimalik mitteparameetriline hinnang mediaanile.

1.6 Ülesanded

1. Kui eeldame, et vaatlused X on normaaljaotusega, siis millise testi abil kontrollime hüpoteesi $H_0 : \text{med}(X) = 3$?

2. Meil on väike valim, mille uurimiseks sooviksime kasutada t-testi. Üheks t-testi eelduseks on nõue, et vaatlused peavad olema normaaljaotusega. Normaaljaotuse eelduse kontrollimiseks teeb statistik mr. Greenhorn formaalse testi (näiteks Shapiro-Wilk'i testi), mis kontrollib nullhüpoteesi H_0 : "vaatlused on normaaljaotusega". Testi olulisustõenäosus on 0,7, millest mr. Greenhorn järeldab, et tema vaatlused on normaaljaotusega ja t-testi kasutamine on õigustatud. Mida peaks mr. Greenhorn tegelikult lisaks tegema, et vastav väide oleks aktsepteeritav?

3. Tehti 8 mõõtmist ja saadi järgmine valim: 10, 11, 12, 15, 20, 30, 55, 102. Testi märgitesti abil, kas uuritava tunnuse mediaan võiks olla 10,5? Milline on olulisustõenäosus? Millise otsuseni jõuad, kui oleksid kasutanud olulisuse nivood 0,05 (aga olulisuse nivool 0,1)? Milline oleks 0,93-usaldusintervall mediaanile?

4. Leia 78,6%-usaldusintervall 0,25-kvantiilile eelmise ülesande andmete põhjal!