Peatükk 7

Tihedusfunktsiooni ja jaotusfunktsiooni hindamine

7.1 Jaotusfunktsiooni hindamine

Jaotusfunktsiooni saab hinnata kasutades empiirilist jaotusfunktsiooni,

$$\begin{aligned}\hat{F}(x) &= \#\{x_i \leq x\}/n \\ &= 1/n \cdot \sum_{i=1}^{n} I(x_i \leq x),\end{aligned}$$

kus I tähistab indikaatorfunktsiooni.

Valimi 1, 4, 6, 9 jaoks leitud parameetriline (leitud normaaljaotuse eeldust kasutades) ja mitteparameetriline jaotusfunktsiooni hinnang on toodud joonisel 7.1.

Empiiriline jaotusfunktsioon on suurepäraste omadustega. Tegemist on nihketa hinnanguga (sest suhteline sagedus on nihketa hinnang tõenäosusele):

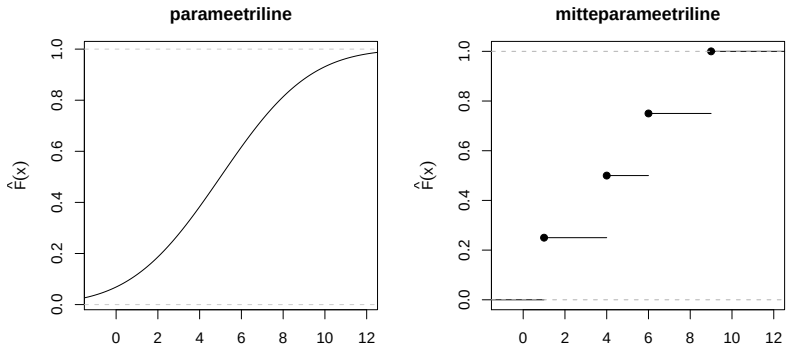
$$\begin{aligned}E\hat{F}(x) &= E\left(1/n \cdot \sum_{i=1}^{n} I(x_i \leq x)\right) \\ &= 1/n \cdot \sum_{i=1}^{n} E\left(I(x_i \leq x)\right) \\ &= 1/n \cdot \sum_{i=1}^{n} F(x) \\ &= F(x)\end{aligned}$$

ja empiiriline jaotusfunktsioon osutub ka mõjusaks hinnanguks (sõltumatute vaatluste korral):

$$\begin{aligned} D\hat{F}(x) &= D\left(1/n \cdot \sum_{i=1}^{n} I(x_i \leq x)\right) \\ &= 1/n^2 \cdot \sum_{i=1}^{n} D\left(I(x_i \leq x)\right) \\ &= 1/n^2 \cdot \sum_{i=1}^{n} F(x)(1-F(x)) \\ &= \frac{F(x)(1-F(x))}{n} \end{aligned}$$

saadud avaldis koondub aga üsna ilmselt nulliks kui $n \to \infty$. Järelikult koondub empiiriline jaotusfunktsioon valimi kasvades tegelikuks jaotusfunktsiooniks (tegemist on mõjusa hinnanguga).

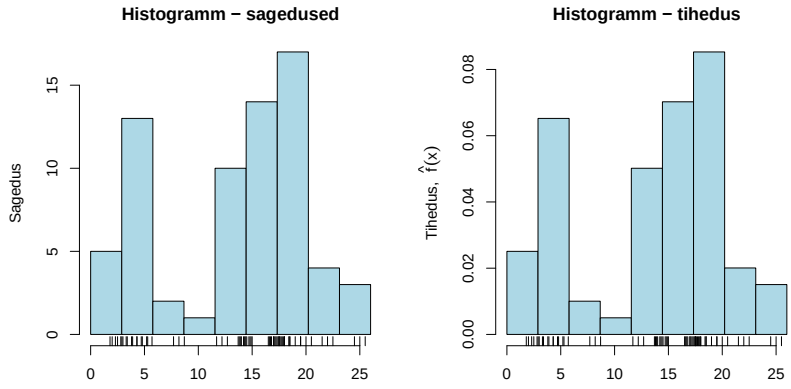
Joonis 7.1: Jaotusfunktsiooni hinnang



7.2 Tihedusfunktsiooni hindamine

Tihedusfunktsiooni mitteparameetrilise hindamisega on mingil kujul arvatavasti iga andmeanalüüsiga tegelenud inimene kokku puutunud — näiteks uurides histogrammi. Tõsi, sageli joonistatakse histogramm küll selliselt, et graafiku y-teljel on mingisse vahemikku sattunud vaatluste arv. Tihedusfunktsiooni hinnangu saamiseks peaksime aga muutma y-telge selliselt, et histogrammi alune pindala oleks üks. Kui kõik histogrammi tulbad on sama laiad, laiusega Δ, ja i. vahemikus on y_i vaatlust, siis tavalise, sagedusi kasutava histogrammi alune pindala on $\sum_i \Delta \cdot y_i = \Delta \cdot \sum_i y_i = \Delta \cdot n$, kus n on vaatluste koguarv. Histogrammist tihedusfunktsiooni hinnangu saamiseks tuleks lihtsalt vaadeldud sagedusi jagada vaatluste arvu ja vahemiku pikkuse korrutisega, vaata ka joonist 7.2.

Joonis 7.2: Histogramm - lihtne võimalus hinnata tihedusfunktsiooni



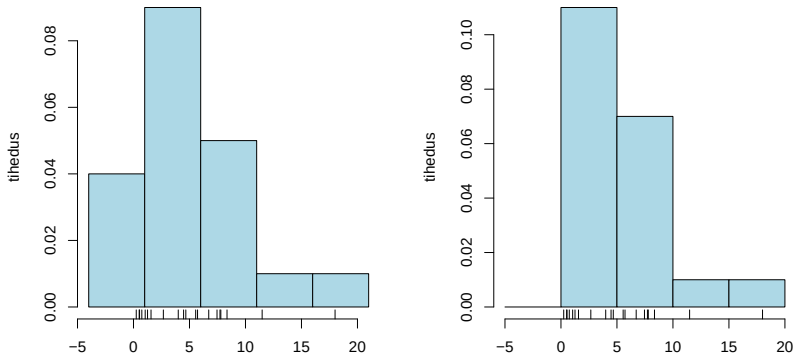
Histogrammi joonistamisel tekivad aga paar raskest küsimust. Kuhu tõmmata tulpade vahelised piirid (vahemike asukohad)? Kui palju vahemikke kasutada (millise laiusega tulpe kasutada)? Vastustest nendele küsimustele võib graafik (ja tema interpretatsioon) sõltuda märkimisväärselt! Vaata ka jooniseid 7.3 ja 7.4.

Lisaks neile kahele küsimusele (vahemike alguspunkti ja laiuse valik) võime muret tunda ka selle üle, kas ehk ei eksisteeri veel mõni teine — loodetavasti täpsem — võimalus hinnata tihedusfunktsiooni. Vaatamegi, kas oskame välja pakkuda mõnda alternatiivset võimalust tihedusfunktsiooni hindamiseks.

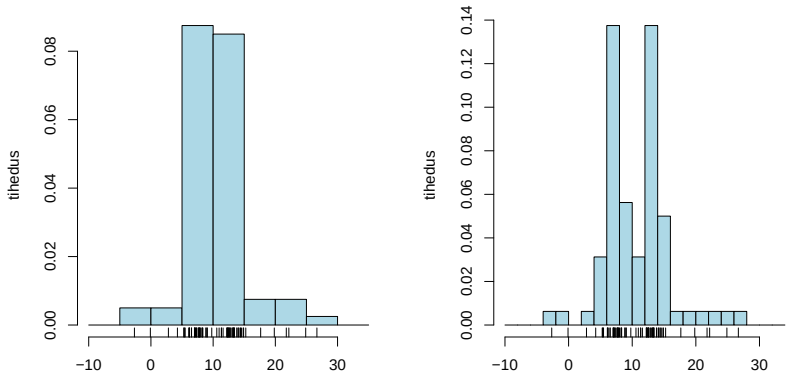
7.2.1 Tihedusfunktsiooni hindamine empiirilise jaotusfunktsiooni abil

Jaotusfunktsioonile on olemas heade omadustega (nihketa, mõjus) hinnang — empiiriline jaotusfunktsioon. Kuna jaotusfunktsiooni teades saab leida tihedusfunktsiooni, $F'(x) = f(x)$, siis äkki on võimalik kasutada empiirilist jaotusfunktsiooni tihedusfunktsiooni hindamiseks? Empiirilise jaotusfunktsiooni tuletis meile midagi väga mõistlikku ei anna (milline on treppfunktsiooni tuletis?). Tuletame aga meelde, kuidas numbrilisi meetodeid kasutades saab hinnata/lähendada funktsiooni tuletist — funktsiooni muutumise kii-

Joonis 7.3: Samad andmed — erinev vahemiku alguspunkt



Joonis 7.4: Samad andmed — erinev tulba laius



rust — punktis x:

$$\hat{F}'(x) = \frac{F(x+h) - F(x-h)}{2h}.$$

Siit võiks tulla mõttele kasutada tihedusfunktsiooni hindamiseks kohas x avaldist:

$$\hat{f}(x) = \frac{\hat{F}(x+h) - \hat{F}(x-h)}{2h}.$$

Saadud valemi paremaks mõistmiseks paneme tähele, et $\hat{F}(x+h) - \hat{F}(x-h)$ on vahemikku $(x-h \ldots x+h]$ sattunud vaatluste arv jagatud valimi suurusega (ehk antud vahemikku sattunud vaatluste osakaal kõigist vaatlustest). Tihedusfunktsiooni hinnangu mingis kohas oleks seega seda kohta ümbritsevasse vahemikku sattunud vaatluste arv jagatud valimi suuruse ja vahemiku laiuse korrutisega. Saadud arvutuseeskiri on väga sarnane sellele, milleni jõudsime histogrammi abil tihedusfunktsiooni hinnates. Kui võrdleksime empiirilist jaotusfunktsiooni kasutavat hinnangut (kasutades vahemikku laiusega $2h$) ja histogrammi abil saadud hinnangut (histogrammi tulbad laiusega $\Delta = 2h$), siis selgub, et histogrammi tulpade keskpunktides langevad saadud hinnangud alati kokku. Võluv on see, et kokkulangemine toimub alati, ükskõik millise alguspunkti me histogrammi vahemikele ka ei valiks. Seega oleme leidnud meetodi, mille puhul ei kerki enam üles küsimust vahemike alguspunktide paigutamise kohta. Üks lihtne andmestik ja selle andmestiku põhjal empiirilist jaotusfunktsiooni kasutades saadud hinnang tihedusfunktsioonile on toodud pildil 7.5.

Saadud tihedusfunktsiooni hinnangu võime soovi korral kirjutada veidi teisele kujule:

$$\begin{aligned}
\hat{f}(x) &= \frac{\hat{F}(x+h) - \hat{F}(x-h)}{2h} \\
&= \frac{\sum_{i=1}^{n} I(x_i - h < x \leq x_i + h)/n}{2h} \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{I(x_i - h < x \leq x_i + h)}{2h}.
\end{aligned}$$

Aga

$$\frac{I(x_i - h < x \leq x_i + h)}{2h}$$

on ühtlase jaotuse $U(x_i - h, x_i + h)$ tihedusfunktsioon, seega on saadud tihedusfunktsiooni hinnang vaadeldav kui n erinevas kohas paikneva ühtlase jaotuse tihedusfunktsiooni keskmine. Tulemuseks on loomulikult tihedusfunktsioon, sest kõik funktsioonid, mille keskmist me leiame, on mittenegatiivsed

ja nende alune pindala on 1, seega on ka samade omadustega saadud "keskmine". Samas võiks ju kasutada ühtlase jaotuse tihedusfunktsioonide asemel mõnda teist tihedusfunktsiooni, näiteks võiksime leida tihedusfunktsiooni kui tiheduste $N(x_1, h)$, $N(x_2, h)$ jne keskmise. Juhul, kui kasutame ühtlase jaotuse asemel vaatluste x_i ümber tsentreeritud normaaljaotuseid, saaksime varem kasutatud andmete (pilt 7.5) põhjal uue hinnangu tihedusfunktsioonile, vaata pilti 7.6. Sellist lähenemist — tihedusfunktsiooni hindamist paljude tihedusfunktsioonide keskmise abil — uurime järgnevas alampeatükis veidi põhjalikumalt.

7.2.2 Tuumameetod tihedusfunktsiooni hindamiseks

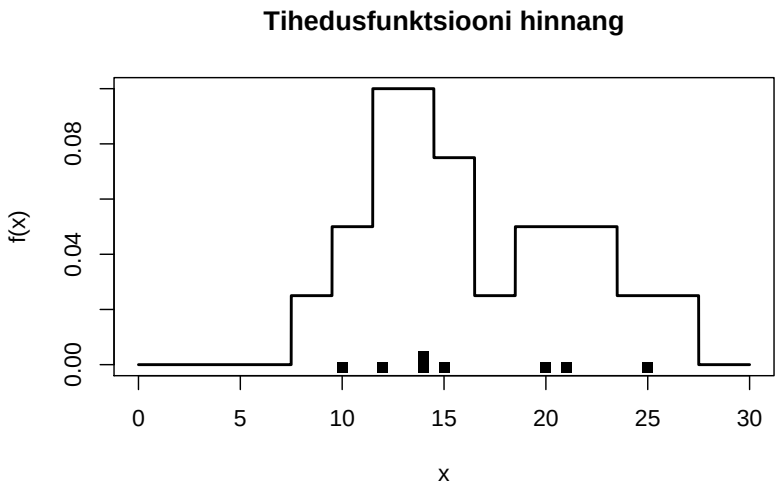
Olgu $K(x)$ mingi keskväärtusega 0 ja lõpliku dispersiooniga juhusliku suuruse tihedusfunktsioon. Vahel nõutakse täiendavalt, et tegemist oleks sümmeetrilise jaotusega, $K(x) = K(-x)$.

Siis funktsiooni

$$\hat{f}(x, h) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - x_i}{h}\right)$$

kutsutakse tuumameetodi abil leitud tihedusfunktsiooni hinnanguks (*Kernel density estimator, KDE*). Tihedusfunktsiooni hindamisel kasutatud tihe-

Joonis 7.5: Empiriilist jaotusfunktsiooni kasutav tihedusfunktsiooni hinnang



dusfunktsiooni $K(x)$ nimetatakse tuumafunktsiooniks (*Kernel function*).

Valemi selgituseks: Olgu antud juhuslik suurus $X \sim K(x)$, keskväärtusega 0 ja mingi standardhälbega (näiteks $\sigma = 1$). Siis $Y := hX + \mu$ oleks keskväärtusega μ ja standardhälbega h (või $h\sigma$, kui $\sigma \neq 1$) juhuslik suurus. Aga juhusliku suuruse Y tihedusfunktsioon $f_Y(y)$ avaldub kujul

$$f_Y(y) = \frac{1}{h} K\left(\frac{y-\mu}{h}\right). \tag{7.1}$$

Selles veendumiseks võid meelde tuletada, et juhusliku suuruse X funktsiooni $Y = g(X)$ tihedusfunktsioon avaldub kujul

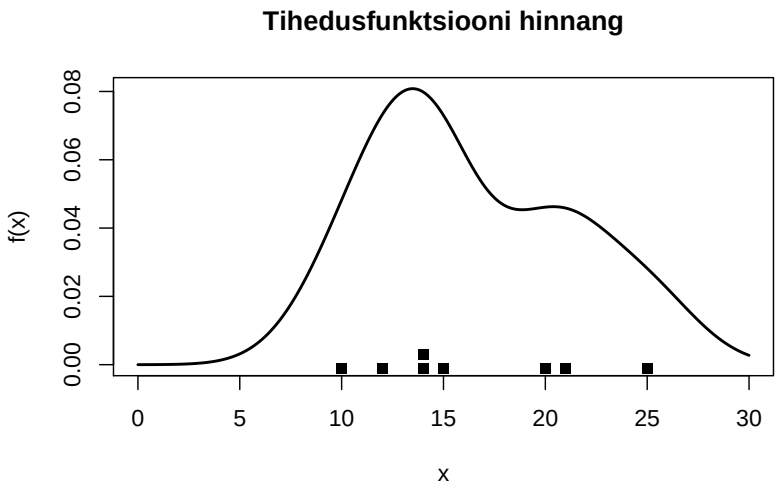
$$f_Y(y) = \left|\frac{1}{g'(g^{-1}(y))}\right| \cdot f_X(g^{-1}(y)).$$

Praegusel juhul aga $g(x) = hx + \mu$, $g'(x) = h$ ja $g^{-1}(y) = (y-\mu)/h$. Seega saamegi tulemuseks valemi 7.1.

Kui on antud keskväärtusega 0 ja dispersiooniga σ juhuslik suurus $X \sim K(x)$, siis $f_Y(x) = \frac{1}{h} K\left(\frac{x-x_i}{h}\right)$ on keskväärtusega x_i (ja standardhälbega $h\sigma$) juhusliku suuruse tihedusfunktsioon.

Tuumahinnangu saamiseks tihedusfunktsioonile võtame seega mingi etteantud hajuvusega $h\sigma$ tihedusfunktsioone, tõstame nad olemasolevate vaat-

Joonis 7.6: Tihedusfunktsiooni hinnang kui paljude normaaljaotuste keskmine



luste kohale (nii et i. tiheduse keskväärtus oleks x_i), leiame saadud tihedusfunktsioonide summa ja jagame vaatluste arvuga (et tulemuseks saadud funktsiooni graafiku alune pindala ikka 1 tuleks, nagu tihedusfunktsioonile kohane).

Millise tulemuse saame, jääb muidugi mingil määral sõltuma nii valitud tuumafunktsioonist kui ka kasutatud h väärtusest. Joonisel 7.7 on näha paari enamkasutatavat tuumafunktsiooni, joonisel 7.8 võib aga näha, milliste tihedusfunktsiooni hinnanguteni (sama andmestikku kasutades) jõuame erinevate tuumafunktsioonide korral. Sagedamini kasutamist leidvad tuumafunktsioonid on

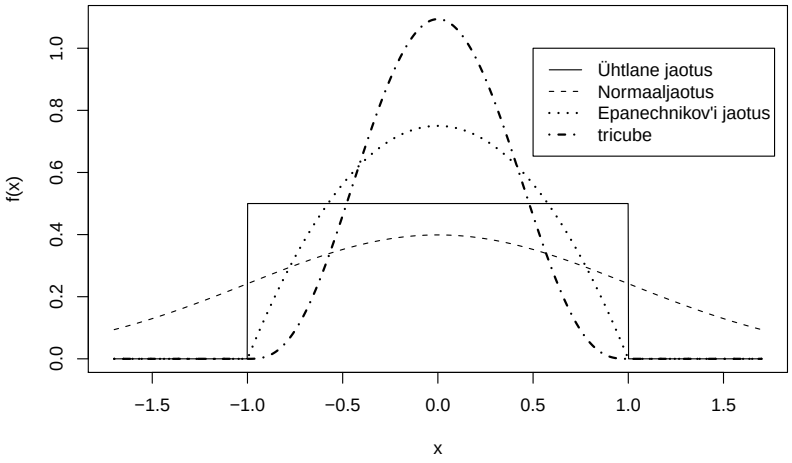
- Ühtlane jaotus $U(-1,1)$:
$$K(x) = 0,5I(|x| \leq 1)$$
- Normaaljaotus $N(0,1)$:
$$K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$$
- Epanechnikov'i jaotus:
$$K(x) = \frac{3}{4}(1-x^2)I(|x| \leq 1)$$
- Tricube:
$$K(x) = \frac{35}{32}(1-x^2)^3 I(|x| \leq 1).$$

Normaaljaotuse tihedusfunktsioon ei muutu kunagi päris nulliks — seega saadakse normaaljaotust tuumafunktsioonina kasutades selline tihedusfunktsiooni hinnang, kus ka vägaväga suurte ja väga-väga väikeste väärtuste saamine on põhimõtteliselt võimalik (kuigi ebatõenäoline). Seevastu Epanechnikovi, tricube ja ühtlase jaotuse puhul ei mõjuta kaugemal kui h ühikut paiknevad vaatlused tihedusfunktsiooni hinnangut antud punktis (muul moel kui valimi suuruse kaudu).

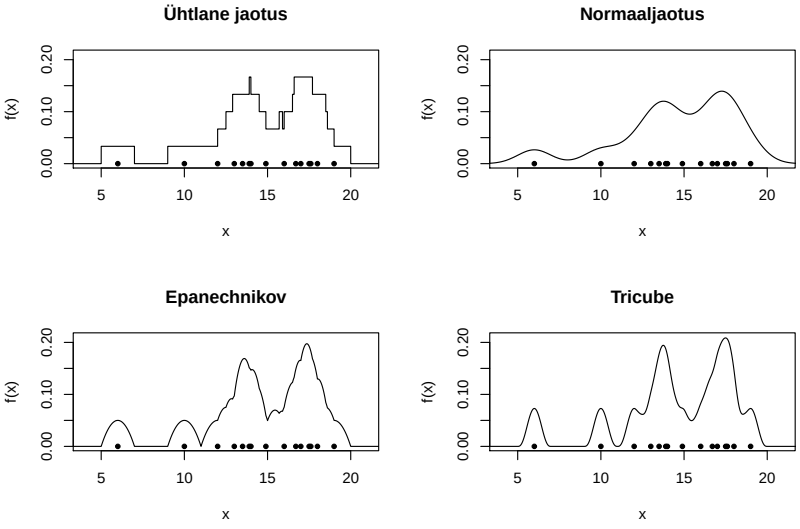
Paraku jääb hetkel veel lahtiseks küsimus, kui suure hajuvusega jaotust — või kui suurt "akna" laiust h — peaksime kasutama (olgu ta siis ühtlane, normaaljaotus või midagi muud). Tulemused võivad tulla üsnagi erinevad, sõltuvalt tehtud valikust. Vaata ka pilti 7.9. Siludes liiga palju — valides suure h-i (histogrammi puhul Δ) väärtuse — saame küll väikese dispersiooniga hinnangu (paljude binaarsete — x_i kuulub antud vahemikku või ei —

vaatluse keskmine annab üsna täpse hinnangu antud vahemikku sattumise tõenäosusele) aga see-eest võib märkamatuks jääda tegeliku tihedusfunkt-

Joonis 7.7: Sagedamini kasutatavaid tuumafunktsioone

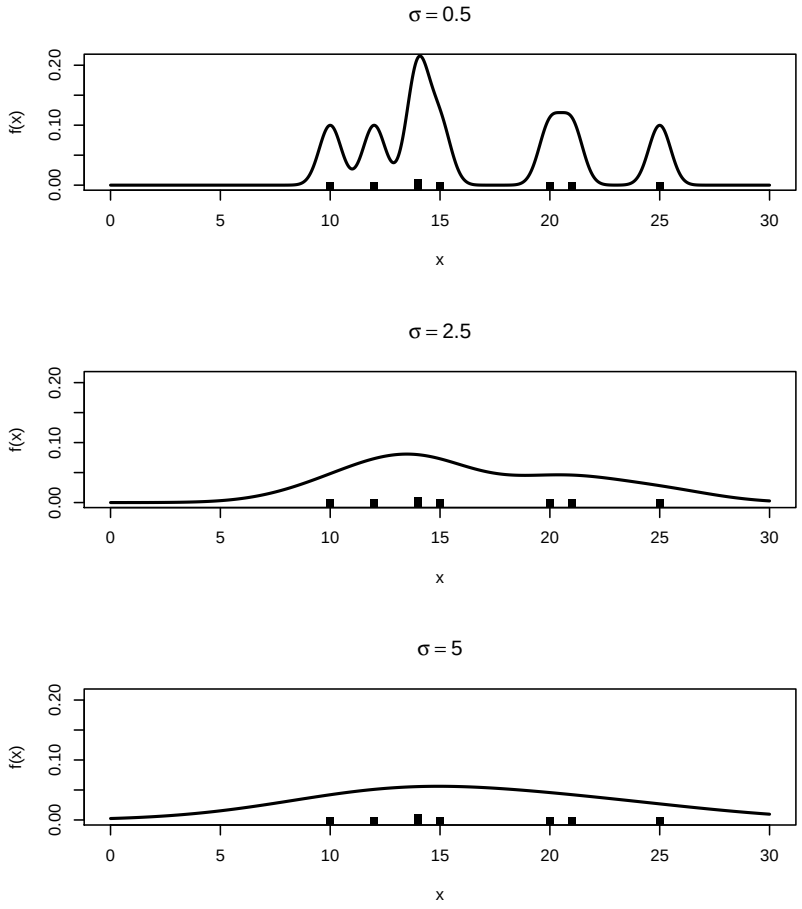


Joonis 7.8: Erinevaid tuumafunktsioone kasutavad tihedusfunktsiooni hinnangud



siooni mõni tipp või lohk. Liiga väikese h puhul on meil põhimõtteliselt võimalik saada küll kiiremini muutuvat tihedusfunktsiooni hinnangut, mis võiks ju sobituda paremini tegeliku tihedusfunktsiooniga, aga paraku läheb suureks ka valimi juhuslikkusest tingitud hinnanguviga.

Joonis 7.9: Tihedusfunktsiooni hinnang kui paljude normaaljaotuste keskmine



Sobivat akna laiust h võiksime otsida selliselt, et teda kasutades saadud hinnang $\hat{f}(x)$ oleks võimalikult täpne, ehk ruutviga $(\hat{f}(x) - f(x))^2$ oleks võimalikult väike. Praku on hinnangu viga iga x väärtuse korral erinev, meie sooviksime aga kogu erinevust iseloomustada ühe numbri abil — siis saaksime seda tekkivat "üldist" viga ka minimiseerida. Selleks üheks üldiseks näitajaks

võiks sobida näiteks integreeritud ruutviga (*ISE - Integrated Squared Error*):

$$ISE = \int (\hat{f}(x) - f(x))^2 dx$$

või keskmine integreeritud ruutviga (MISE), mille puhul minimiseeritakse keskmist ISE't üle kõikmõeldavate valimite:

$$MISE = \mathrm{E} \int (\hat{f}(x) - f(x))^2 dx.$$

Mainitud statistikute (ISE ja MISE) arvutamisega jääme praktikas muidugi hätta, sest nende leidmiseks peaksime teadma tegelikku tihedusfunktsiooni $f(x)$. Küll aga saab neid näitajaid kasutada teoreetiliste tulemuste leidmiseks (Näiteks: kui suurendame valimi mahtu n, siis MISE väheneb kõige kiiremini, kui valime akna laiuse — ehk tuumafunktsiooni hajuvuse — järgmise eeskirja järgi: $h = cn^{-1/5}$, kus c on mingi hinnatavast tihedusfunktsioonist ja tuumafunktsioonist sõltuv konstant...).

Õnneks selgub siiski, et mõlemad näitajad on (suhteliselt) lihtsasti hinnatavad — konstandi täpsuseni küll. Nimelt:

$$\begin{aligned} ISE &= \int (\hat{f}(x) - f(x))^2 dx \\ &= \int \hat{f}(x)^2 dx - 2\int \hat{f}(x) \cdot f(x) dx + \int f(x)^2 dx \\ &= \int \hat{f}(x)^2 dx - 2\mathrm{E}\hat{f}(x) + \int f(x)^2 dx \\ &\approx \int \hat{f}(x)^2 dx - 2\overline{\hat{f}(x)} + \int f(x)^2 dx. \end{aligned}$$

Kui kaalume, millist aknalaiust h kasutada või millist tuumafunktsiooni $K(x)$ eelistada, siis viimane liidetavatest ei muutu — viimane liidetavatest ei sõltu ju üldse hinnangust $\hat{f}(x)$. Seega võime ta minimiseerimisülesandest eemaldada. Ülejäänud on aga kõik arvutatav — muidugi kui asendame keskväärtuse valimi keskmisega, $\mathrm{E}\hat{f}(x) \approx \overline{\hat{f}(x)}$.

Tõsi küll, nihketa hinnangu saamiseks $\mathrm{E}\hat{f}(x)$-le tuleks valimi keskmine seekord leida veidi ebatavalisel teel. Nimelt tähistame $\hat{f}_{-i}$-ga tihedusfunktsiooni hinnangu, mille saame ilma i. vaatlust kasutamata (eemaldame valimist i. vaatluse ja leiame siis tihedusfunktsiooni hinnangu). Siis $\overline{\hat{f}(x)} = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i)$ on nihketa hinnanguks $\mathrm{E}\hat{f}(x)$-le. Saadud meetodit — vali h vastavalt eeskirjale

$$h = \arg\min \int \hat{f}(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i)$$

tuntakse nihketa ristvalideerimise nime all (*unbiased cross-validation*).

7.3 Hinnangu teoreetilised omadused

Järgnevalt toome ära mõned teoreetilised tulemused, mis aitavad välja tuua tuumameetodi eeliseid. Tõestuskäike ja detaile vaata Wasserman (2006).

Oletame, et soovime minimiseerida keskmist integreeritud ruutviga (MISE). Siis saab teatud matemaatilistel eeldustel (tihedusfunktsiooni tuletis on pidev jt) näidata, et histogrammi tulba optimaalne laius (tulba laius mille puhul keskmine integreeritud ruutviga üle võimalike valimite tuleb kõige väiksem) on kujul $h_{opt} = n^{-1/3} \cdot c_1$. Kui histogrammi tegemisel kasutada optimaalset tulba laiust, siis $MISE \sim c_2 n^{-2/3}$. Kordajad c_1 ja c_2 on teatavad konstandid, mis sõltuvad hinnatavast tihedusfunktsioonist endast.

Kui aga kasutada tuumameetodit tihedusfunktsiooni hindamiseks, siis $h_{opt} = n^{-1/5} \cdot c_3$ ja optimaalset "venitajat" kasutatades saavutatakse $MISE \sim c_4 n^{-4/5}$, kus c_3 ja c_4 on tuumafunktsioonist ja tegelikust tihedusfunktsioonist sõltuvad konstandid (taas on kasutatud teatud matemaatilisi eelduseid, näiteks peab tegeliku tihedusfunktsiooni teine tuletis olema pidev jne).

Miks mainitud teoreetilised tulemused on tähtsad? Kuna $n^{-4/5}$ koondub nulliks kiiremini kui $n^{-2/3}$, siis vähemalt mingist valimi suurusest alates on tuumafunktsioonil baseeruv hinnang täpsem kui histogrammil baseeruv tihedusfunktsiooni hinnang.

Stone'i teoreem väidab, et asümptootiliselt on ristvalideerimise abil saadud aknalaius sama hea (viib sama täpsete hinnanguteni ISE-mõttes) kui optimaalne aknalaius:

Teoreem 7.1 (Stone) *Olgu f tõkestatud. Olgu $\hat{f}_h$ aknalaiust h kasutades saadud hinnang tihedusfunktsioonile. Tähistame ristvalideerimise abil saadud optimaalset aknalaiust h_*-ga. Siis*

$$\frac{\int \left(f(x) - \hat{f}_{h_*}\right)^2 dx}{\inf_h \int \left(f(x) - \hat{f}_h\right)^2 dx} \xrightarrow[a.s.]{} 1.$$

Lisaks saab tõestada tulemuse, et kui informatsiooni tegeliku tihedusfunktsiooni kohta tõepoolest napib, siis pole võimalik leida hinnangut tihedusfunktsioonile, mille MISE (keskmine integreeritud ruutviga) koonduks nulliks kiiremini kui $n^{-4/5}$ (see ei tähenda otseselt, et tuumameetodil saadud hinnang oleks täpseim võimalik, aga see tähendab, et ühelgi hinnangul — kui me just ei tea lisainformatsiooni tegeliku tihedusfunktsiooni kohta — ei

õnnestu hinnangu viga vähendada kiiremini kui tuumahinnangul). Tõsi, kui nõustume, et tihedusfunktsiooni hinnang ise ei pruugi olla tihedusfunktsioon (lubame tal omandada ka negatiivseid väärtuseid), on võimalik leida veelgi kiiremini koonduvaid hinnanguid.

Põhimõtteliselt saab leida ka parima tuumafunktsiooni, mille puhul MISE (vähemalt asümptootiliselt) kõige väiksem oleks. Üheks optimaalseks tuumafunktsiooniks on Epanechnikovi tuumafunktsioon. Keskmine võit täpsuses võrreldes teiste tuumafunktsiooni valikutega on aga tühine, tuumafunktsiooni valikut ei peeta üldiselt eriti kriitiliseks probleemiks. Pigem valitakse tuumafunktsioon antud rakenduse jaoks vajalike omaduste järgi (näiteks: kui kasutame normaaljaotust tuumafunktsioonina, siis ei muutu saadud tihedusfunktsiooni hinnang kuskil matemaatilises mõttes nulliks — erinevalt näiteks Epanechnikovi või ühtlasest jaotusest tuleneva tuumafunktsiooni abil saadud hinnangutest. Sõltuvalt ülesandest võib see olla soovitud või ebasoovitav tulemus).