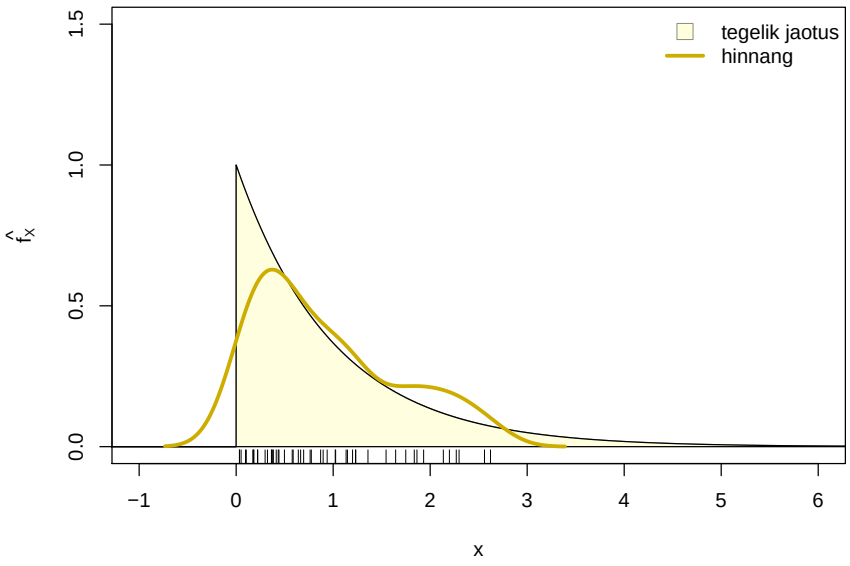
7.4 Lisainformatsiooni kasutamisest

Arvuti teab vaid seda informatsiooni, mis tema käsutusse on antud. Vahel arvatakse ekslikult, et piisab statistikapaketile vaid kogutud andmete edastamisest. Enamasti on sellise lähenemise puhul tulemuseks kehvapoolsed tulemused. Kui uurija viitsib mõista, millist tunnust ta uurib, siis avastab ta sageli, et teab tunnuse käitumise kohta märkimisväärselt palju taustainformatsiooni. Näiteks kui uuritavaks tunnuseks on kas kaal või vanus, siis peavad uuritava tunnuse väärtused olema positiivsed, $X > 0$. Kui me arvutiga seda taustainformatsiooni aga ei jaga, hinnatakse andmete pealt sageli tihedusfunktsioon, mis justnagu lubaks ka negatiivsete vaatluste tekkimist, vaata näiteks joonist 7.10.

Joonis 7.10: Tihedusfunktsiooni hinnang ilma lisainformatsiooni kasutamata



Uuritava tunnuse kohta olevat taustainformatsiooni tuleks parimate tulemuste saamiseks ka kasutada. Kuidas aga tihedusfunktsiooni hindamisel kasutada teadmist, et uuritava tunnuse väärtused ei saa olla negatiivsed?

Toome siinkohal ära kaks erinevat võimalust lisainformatsiooni $X > 0$ kasutamiseks.

Tunnuse transformeerimine

Kui $X \in (0 \ldots \infty)$, siis $\ln(X) \in (-\infty \ldots \infty)$ ehk juhuslik suurus $\ln(X)$ võib paikneda juba reaalttelje mistahes piirkonnas (tema paiknemise kohta meil enam lisainformatsiooni pole). Sageli hinnataksegi juhusliku suuruse $Y := \ln(X)$ tihedusfunktsioon ja leitakse seejärel juba matemaatiliste teisenduste abil, milline on juhusliku suuruse $X = \exp(Y)$ tihedusfunktsioon.

Meeldetuletuseks: Kui soovime leida juhusliku suuruse Y funktsiooni $X = g(Y)$ tihedusfunktsiooni, siis saame vajaliku arvutuse teha järgmise valemi abil (eeldame, et g on funktsiooni määramispiirkonnas I pidev ja monotoonne):

$$f_X(x) = \begin{cases} f_Y(g^{-1}(x)) \cdot \left|\left(g^{-1}\right)'(x)\right|, & x \in I \\ 0 & x \notin I \end{cases} .$$

Kui kasutasime transformatsiooni $X = \exp(Y)$, siis $g(x) = \exp(x)$, $g^{-1}(x) = \ln(x)$ ja $(g^{-1})'(x) = 1/x$. Seega $f_X(x) = f_{\ln(X)}(\ln(x))/x$. Saadud valem võimaldab minna tunnuse logaritmitud väärtustele leitud tihedusfunktsioonilt üle mittetransformeeritud juhuslike suuruste tihedusele. Antud meetodi rakendamisel saadud tulemust iseloomustab joonis 7.11.

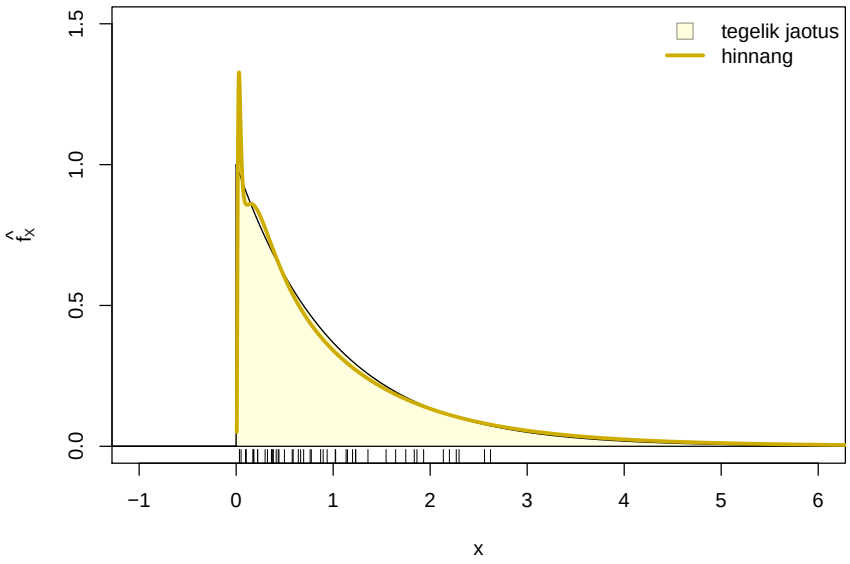
Vaatluste peegeldamine

Teine meetod mida kasutatakse kui $X > 0$ kasutab vaatluste peegeldamist nullpunkti suhtes (ja peegeldatud vaatluste lisamises esialgsele valimile). Kui mõõdetud uuritava tunnuse väärtused on näiteks 12, 15, 16, ... siis lisatakse täiendavalt valimile veel vaatlused -12, -15, -16, ... Saadud uue valimi (mis on esialgsest valimist 2 korda suurem) põhjal leitakse esialgne hinnang tihedusfunktsioonile $\hat{f}_1$. Saadud proto-tihedusfunktsioonist täpselt pool paikneb nüüd negatiivsete väärtuste poolel (juhul, kui kasutame sümmeetrilist tuumafunktsiooni). Tegelikult muidugi ei saa meid huvitav juhuslik suurus X omandada negatiivseid väärtuseid. Seega peab $f_X(x) = 0$, kui $x \leq 0$. Kui võtame ette peegeldatud väärtuseid kasutades saadud proto-tihedusfunktsiooni hinnangu $\hat{f}_1$ ja muudame ta nullist väiksemate x väärtuste korral võrdseks nulliga, siis peame ülejäänud, nullist suuremat poolt tõstma 2 korda (et tulemuseks saadud funktsioon ikkagi tihedusfunktsioon oleks):

$$\hat{f}_X(x) = \begin{cases} 0, & x \leq 0 \\ 2\hat{f}_1, & x > 0 \end{cases} .$$

Vaatluste peegeldamise ehk valimi kahekordistamise meetodil saadud hinnang tihedusfunktsioonile on toodud joonisel 7.12.

Joonis 7.11: Tihedusfunktsiooni hinnang ($X > 0$) kasutades logaritmilist transformatsiooni

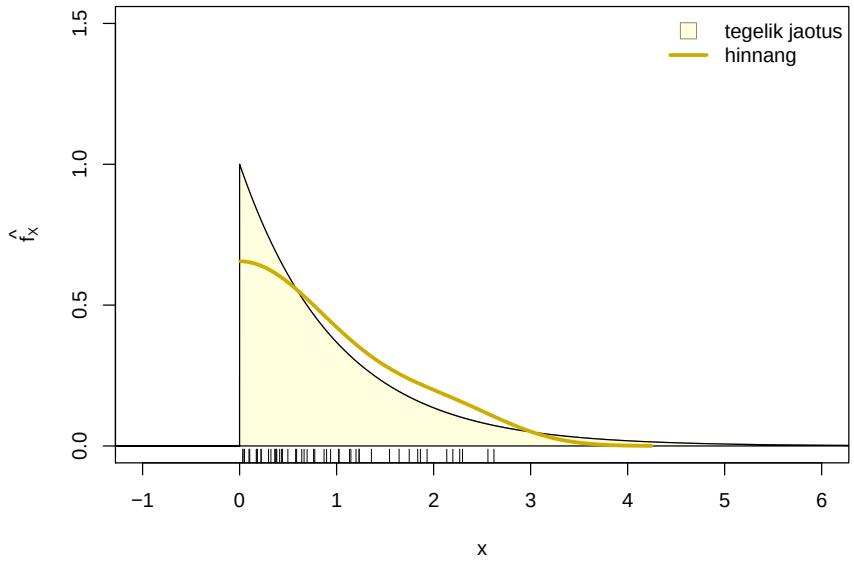


Kumb kahest väljapakutud meetodist paremini töötab, sõltub konkreetsest andmestikust. Logaritmilist transformatsiooni kasutava meetodi puhul on tihedusfunktsiooni hinnangu parempoolne piirväärtus nullpunktis alati 0 ($\lim_{x\to 0+} \hat{f}_X(x) = 0$) ehk saadud tihedusfunktsiooni hinnang on pidev punkti $x = 0$ ümbruses. Seevastu väärtuste peegeldamise meetodi puhul sellist nõuet pole, tihedusfunktsiooni hinnangul esineb sageli katkevuspunkt kohas $x = 0$.

7.5 Rakendusnäide

Üheks sageli statistikule ettetulevaks ülesandeks on klassifitseerimisülesanne: vaadeldud tunnuse või tunnuste põhjal tuleb uuritavad objektid jagada kahte või enamasse teadaolevasse klassi (pank soovib kliente jagada laenu tagasimaksvateks ja makseraskustesse sattuvateks klientideks; geneetik soovib mõne välise tunnuse järgi aru saada, kas uuritav on AA, AB või BB genotüübiga; botaanik soovib heinakõrre laiuse või mõne muu tunnuse põhjal määrata mis liiki taimega on tegemist jne). Tavaliselt on sellise ülesande

Joonis 7.12: Tihedusfunktsiooni hinnang ($X > 0$) kasutades valimi kahekordistamise ehk vaatluste peegeldamise meetodit



lahendamiseks kasutada valim, kus eksperdi poolt või kalleid analüüse kasutades on määratud objektide klassikuuluvus. Statistiku ülesandeks on valimi eeskujul lahterdada ka uued objektid etteantud klassidesse.

Jagagem uuritavaid objekte k-sse erinevasse klassi. Oletame, et on teada iga klassi esinemistõenäosus uuritavas populatsioonis (p_1, $p_2, \ldots, p_k$) ja teame ka klassifitseerimiseks kasutatava tunnuse (või tunnuste) jaotust iga klassi jaoks eraldi — teame, milline on tunnuse X jaotus liiki A kuuluvate taimede korral, milline on tunnuse X jaotus liiki B kuuluvate taimede jaoks jne ehk eeldame, et antud on tihedused $f_1 := f_{X|klass=1}, f_2 := f_{X|klass=2}, \ldots, f_k$. Kui peaksime nüüd klassifitseerima uue objekti, mille puhul $X = x$, siis võiksime esmalt leida tinglikud tõenäosused $P(\text{objekt kuulub klassi } 1|X = x)$,...,$P(\text{objekt kuulub klassi } k|X = x)$ kasutades tingliku tõenäosuse valemit:

$$P(\text{objekt kuulub klassi } i|X = x) = \frac{f_i(x) \cdot p_i}{f_X(x)},$$

kus f_X on tunnuse X jaotus uuritavas populatsioonis (üle kõigi klasside):

$$f_X(x) = f_1 p_1 + f_2 p_2 + \ldots + f_k p_k.$$

Kasutades leitud tinglikke tõenäosuseid on aga võimalik uut uuritavat klassifitseerida — näiteks määrates ta kõige tõenäolisemasse klassi.

Loomulikult pole reaalse ülesande korral võimalik kasutada ülesande lahendamiseks tõenäosuseid $p_1, \ldots, p_k$ ega ka tihedusfunktsioone $f_1, \ldots, f_k$, küll aga on võimalik mõlemaid suuruseid hinnata uurija käsutuses oleva valimi põhjal. Saadud hinnanguid kasutades on võimalik leida ka hinnanguid tinglikele tõenäosustele.

Vaata ka joonist 7.13.

7.6 Mitmemõõtmeline tihedus

Ka mitmemõõtmelist tihedusfunktsiooni saab hinnata tuumameetodi abil. Kasutada tuleb lihtsalt mitmemõõtmelist tuumafunktsiooni, mida on näiteks võimalik konstrueerida ühemõõtmeliste tuumafunktsioonide korrutamise teel:

$$K(x, y) = K(x) \cdot K(y).$$

Kasutades ülaltoodud kahemõõtmelist tuumafunktsiooni saame leida kahe tunnuse ühise tihedusfunktsiooni hinnangu:

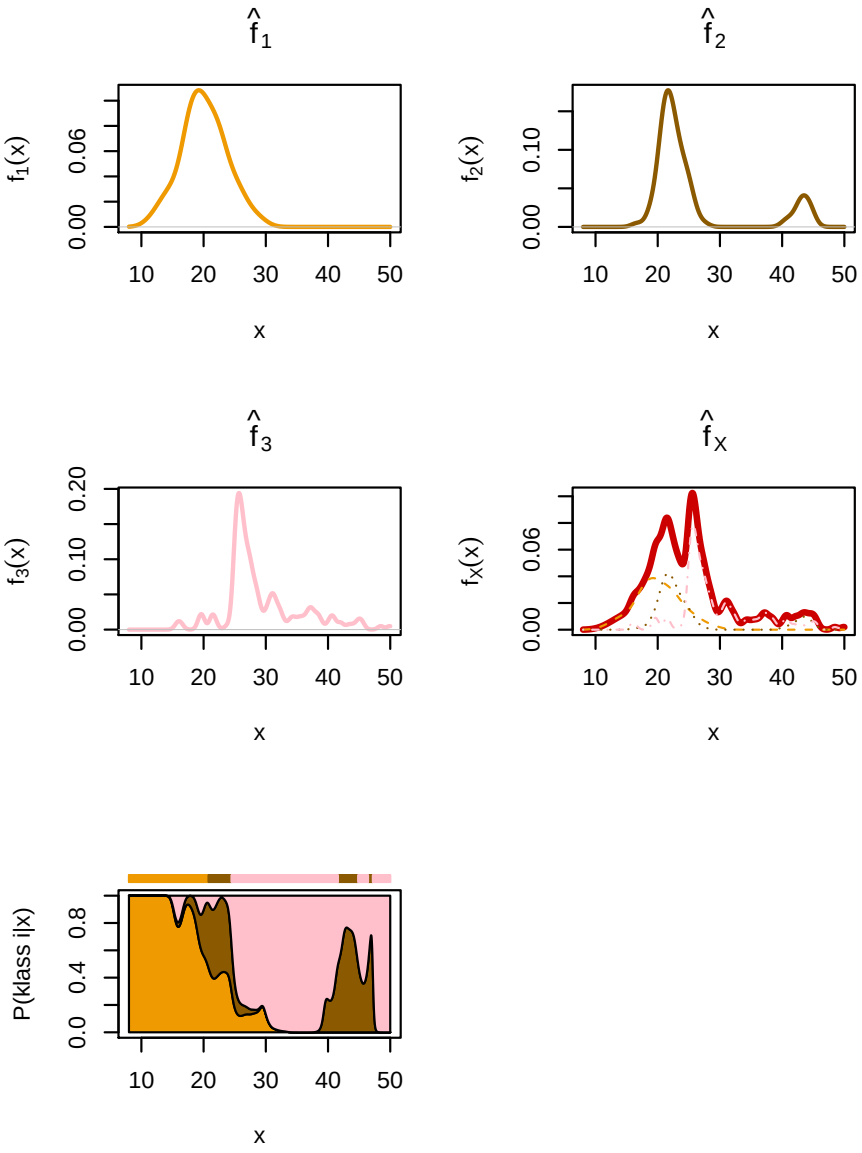
$$\hat{f}_{X,Y}(x, y) = \frac{1}{n} \sum_{i=1}^{n} K_{h_x h_y}(x - x_i, y - y_i) \tag{7.2}$$

$$\hat{f}_{X,Y}(x, y) = \frac{1}{n h_x h_y} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_x}\right) K\left(\frac{y - y_i}{h_y}\right) \tag{7.3}$$

Vaata ka joonist 7.14.

Tuumameetodit kasutatakse harva rohkem kui 2-mõõtmelise tiheduse hindamiseks. Kui uurime 3- või enamamõõtmelisi tihedusfunktsioone tuleb peaaegu alati teha mõistliku täpsuse saavutamiseks mingeid eelduseid tunnuste omavahelise sõltuvusstruktuuri kohta. Ilma selliste lisaeeldusteta jääb tihedusfunktsiooni hinnang tüüpiliste valimi suuruste juures väga ebatäpseks (talutava täpsuse saavutamiseks võime vajada väga suurt valimit). Selliseid sobivaid lisaeeldusi on aga sageli mugavam sisse tuua mõne teise tihedusfunktsiooni hindamiseks mõeldud meetodi korral.

Joonis 7.13: Klassifitseerimisülesande lahendamine tihedusfunktsiooni hinnanguid kasutades



Joonis 7.14: Kahemõõtmeline tuumafunktsioon ja tema abil saadud hinnang kahe tunnuse ühisele tihedusfunktsioonile