Peatükk 3

Juhuslikkuse kirjeldamine

Kuidas iseloomustada midagi täiesti juhuslikku ja ebakindlat
ehk
Tõenäosusest, juhuslikust suurusest ja tema jaotusest

Kuitahes põhjalikult me mõnda looduses aset leidvat protsessi ka ei uuriks, ei suuda me sageli siiski täpselt ette öelda, mis juhtub järgmisel korral. Karumammi võib talvel sünnitada kaks armsat karubeebit, aga võib sünnitada ka ühe (või ei sünnita ühtegi); kahe roosade õitega iirise järglane võib olla valgete õitega, roosade õitega või hoopis punaste õitega; mulda külvatud seemnest võib sirguda sihvakas taim, aga seeme võib ka idanemata jääda.

Kuidas kirjeldada teistele teadlastele, mis juhtub eksperimendi tulemusena, kui juhtuda võib nii või naa? Kirjeldame kõiki võimalusi, mis juhtuda võib? Kirjeldame, et kui jõe ülemjooksu süvendame, siis võivad jõest vähid kaduda aga ei pruugi, ja kui me süvendustöid ette ei võta, ka siis võivad vähid jõest kaduda aga ei pruugi? Sellisest juhuslikkuse kirjeldamisest kindlasti ei piisa tarkade otsuste tegemiseks. Kuidas siis juhuslikkust täpsemalt kirjeldada?

3.1 Suhteline sagedus ja tõenäosus

Mingi sündmuse A suhteline sagedus on sündmuse A toimumiste arv jagatud kõigi katsete (või vaatluste) arvuga.

Näide: Kümnes vähirikkas jões tehti süvendustöid. Süvendatud jõgedest kaheksas kadusid vähid viie aasta jooksul. Sündmuse "Viie aasta jooksul peale süvendustöid kaovad vähid jõest" toimumise suhteline sagedus on $8/10 = 0,8$ ehk 80%.

Paraku on suhtelise sageduse kasutamisel teaduses üks väga tõsine puudus. Nimelt sama nähtust kirjeldades võivad teadlased saada väga erinevaid suhtelisi sagedusi. Näiteks oletame, et kaks teadlast soovivad kirjeldada, kui sageli koorub linnu X munast emaslind. Üks teadlane, Mari (Tartu Ülikoolist), vaatles kolme linnupoja koorumist. Kahest munast koorusid emaslinnud, seega oli emaslinnu koorumise suhteline sagedus Malle jaoks 2/3. Sama linnuliiki uuris ka Yung-Ji (Pekingi 11. Riiklik Ülikool). Temal koorus kolmest linnumunast kõigest üks emaslind, seega oli emaslinnu koorumise suhteline sagedus Yung-Ji jaoks 1/3.

Vähe sellest - hiljem vaatles Mari veel ühe liigist X pärit linnupoja koorumist. Seekord koorus munast isane linnupoeg. Mari arvutas uuesti emaslinnu koorumise suhtelise sageduse ja sai tulemuseks 2/4=0,5.

Mingi sündmuse toimumise sagedust on seega üsna raske teaduslikult kirjeldada kasutades suhtelist sagedust - sest iga teadlane võib saada erineva tulemuse ja saadud numbritest on raske ühte numbrit teisest paremaks pidada.

Mis oleks lahendus? Selgub, et korrates sama katset samades tingimustes, hakkab sündmuse A toimumise suhteline sagedus lähenema mingile numbrile. Vähe sellest — kui keegi teine teadlane kordab sama katset samades tingimustes, siis hakkab ka temal sündmuse A toimumise suhteline sagedus lähenema samale numbrile. Kui mõlemad teadlased saaksid oma katset korrata lõpmatu palju kordi, siis jõuaksid nad ühe ja sama tulemuseni. Seda tulemust — sündmuse A toimumise suhtelist sagedust siis, kui katset korratakse lõpmatu palju kordi, kutsutaksegi sündmuse A toimumise tõenäosuseks (antud katsetingimuste korral). Toimuvat iseloomustab joonis 3.1.

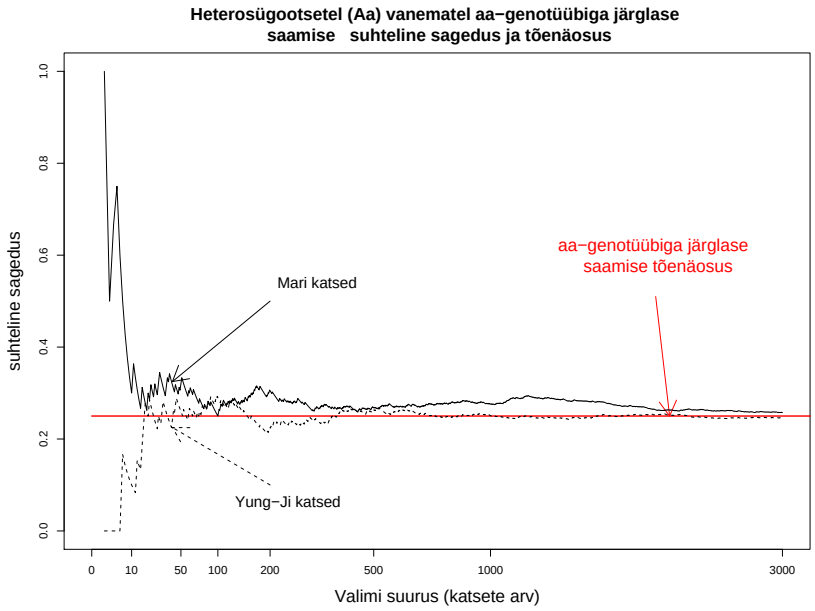
Definitsioon 3.1 *Juhusliku sündmuse toimumise tõenäosuseks $P(A)$ nimetatakse sündmuse A toimumise suhtelist sagedust peale lõpmatult paljude katsete sooritamist. Matemaatiliselt korrektselt kirja pandult:*

$$P(A) = \lim_{n\to\infty} \frac{k}{n}$$

ehk tõenäosus on sündmuse toimumise suhtelise sageduse piirväärtus kui katsete arv läheneb lõpmatusele.

Toodud definitsiooni põhjal saab kohe kirja panna mõned tõenäosuse omadused:

- tõenäosus on alati 0 ja 1 vahel;
- võimatu (mitte kunagi toimuva sündmuse) tõenäosus on 0 $(= 0/\infty)$;



Joonis 3.1: Heterosügootsetel (Aa-genotüübiga) vanematel homosügootse (aa-genotüübiga) järglase saamise tõenäosus ja suhteline sagedus

- alati toimuva sündmuse tõenäosus on 1 ($= \frac{n}{n}$);
- sündmus, mille tõenäosus on 0, võib toimuda ($\frac{1}{\infty} = 0$).

Samas tasub tähele panna, et toodud tõenäosuse definitsiooni on praktiliselt võimatu kasutada tõenäosuse leidmiseks — sest see eeldaks lõpmatult paljude katsete tegemist ehk teisisõnu lõpmatult paljude andmete olemasolu uuritava nähtuse kohta.

3.1.1 Tõenäosuse leidmisest

Kuidas leida meid huvitava sündmuse toimumise tõenäosust? Selleks on paar võimalust.

Esiteks võime teha tõesti väga palju katseid ja oletada, et meid huvitava sündmuse toimumise suhteline sagedus on peaaegu võrdne tõenäosusega (sest katsete arv n on väga suur). Nagu hiljem näeme, on võimalik iseloomustada ka tekkiva vea võimalikku suurust ja kui võimalik viga on väga väike, võime saadud tulemusega leppida. Antud viisil saab tõenäosust leida vaid ligikaudselt — sest enamasti pole võimalik teha lõpmatult palju katseid.

Teine võimalus seisneb arvutusvalemite kasutamises. Teades mõne või mõningate sündmuste toimumise tõenäosuseid, on vahel võimalik arvutada teiste sündmuste toimumise tõenäosuseid. Näiteks teades sündmuse A toimumise tõenäosust $P(A)$ võime leida sündmuse $\overline{A}$ — sündmus A ei toimu — tõenäosuse kasutades valemit $P(\overline{A}) = 1 - P(A)$.

Üheks võimaluseks tõenäosuse leidmiseks on uuride juhuslikkust tekitavat mehhanismi. Näiteks vaadeldes münti võime jõuda otsusele, et mündi serv on liiga kitsuke — sellele õhku visatud münt seisma jääda ei saa — ja mõlemad küljed on täpselt samasugused. Seega peaks tõenäosus ühe külje ülesjäämiseks mündiviskel olema samasuur kui teise külje ülesjäämiseks. Jätkates arutelu samal moel võime lõpuks jõuda järeldusele et nii kirja kui kulli tulemise tõenäosus peab olema $1/2$. Taoline viis tõenäosuste leidmiseks võib töötada üsnagi hästi, kui juhuslikkuse tekkepõhjused on lihtsad ja hästi mõistetavad. On kasutatav õnnemängude korraldamisel või vahest ka füüsikas mõningate nähtuste kirjeldamisel. Bioloogias tuleb sedavõrd hästi ära kirjeldatud süsteeme harva ette, et võiks puhtalt looma või taime kirjelduse põhjal ära öelda, kui suure tõenäosusega meid huvitav sündmus ette tuleb (koer kui liik ja tema elukeskond pole piisavalt hästi kirjeldatud arvutamaks vaid selle kirjelduse põhjal, kui suure tõenäosusega selle liigi esindaja kirjakandjat hammustab). Samas hakatakse elusorganismide molekulaarstruktuuride kirjeldamisel jõudma sedavõrd kaugele, et paljude huvipakkuvate sündmuste tõenäosuseid võib varsti olla võimalik leida näiteks valkude molekulaarstruktuuri uurides.

Arvutusvalemid ja nende eeldused

Sellest, kuidas tõenäosuseid (teiste, teadaolevate tõenäosuste kaudu) arvutada, võib näiteks põhjaliku käsitluse leida A. Jõgi tõenäosusteooria õpikust ja mitmestki teisest eestikeelsest õpikust. Siinses materialis tõenäosuste arvutusvalemitel pikemalt ei peatuta. Välja toome vaid ühe valemi — koolimatemaatikast tuntud tõenäosuse arvutusvalemi.

Paljud tõenäosusega kokkupuutunud inimesed teavad arvutusvalemit

$$P(A) = k/n, \tag{3.1}$$

kus k on sündmuse A jaoks sootsate võimaluste arv (sündmus A toimub) ja n kõigi võimaluste arv. Sageli kipub aga ununema, et antud valem kehtib ainult siis, kui a) kõik n sündmust on võrdvõimalikud — näiteks kõigi täringu külgede ülespoole jäämise tõenäosus on sama; b) kui kõik n sündmust on teineteist välistavad; c) kõiki võimalusi on täpselt n ja mitte rohkem.

Näide 3.1 *Visatakse täringut (võimalikud katsetulemused {1}, {2}, {3}, {4}, {5}, {6}, $n = 6$). Meid huvitab sündmuse A="täringuviske tulemusel saame rohkem kui 4 silma" toimumise tõenäosus ($k = 2$). Antud juhul võime tõenäosuse $P(A)$ leidmiseks kasutada valemit (3.1): $P(A) = 2/6 = 1/3$ (sest kõik visketulemused on võrdvõimalikud).*

Näide 3.2 *Metsas elab 6 jänest ($n = 6$), neist 2 on valgejänesed (Lepus timidus), teised halljänesed (Lepus europaeus). Metsa serva pannakse ülese jäneselõks. Meid huvitab sündmuse A="lõksu langeb valgejänes" toimumise tõenäosus. Miks antud juhul ei tohi kasutada tõenäosuse $P(A)$ arvutamiseks valemit (3.1)? Sest kõik tõenäosused pole võrdvõimalikud — valgejänes on pelglikum kui halljänes, kardab metsast välja tulla — seega sattub ta metsaservale paigutatud lõksu ka harvemini kui halljänes. Sestap pole meid huvitava tõenäosuse $P(A)$ väärtus $2/6\,(= 1/3)$ — $P(A) < 1/3$.*

3.1.2 Juhuslik suurus ja tema jaotus

Vahel soovime kirjeldada, mis siis ikkagi katse tulemuseks võib olla või mida me vaatluse käigus näha võime. Kui katse tulemus pole üheselt ette määratud (tegemist on juhusliku katsega) siis võib katsel olla palju erinevaid tulemusi – katse tulemuseks võib olla juhuslik suurus. Kõigi võimalike katsetulemuste ettelugemisest jääb aga enamasti väheks.

Kuidas oleks võimalik paremini kirjeldada ühte juhuslikku suurust?

Väheste võimalike väärtustega tunnus

Kui katsel on suhteliselt vähe võimalikke tulemusi (juhuslikul suurusel on vähe võimalikke väärtuseid) siis võime anda iga katsetulemuse kohta tema toimumise tõenäosuse.

Näide: Kahe heterosügootse vanema (Aa) ristamisel on juhusliku suuruse X – järglase genotüüp – võimalikud väärtused x ja nende esinemise tõenäosused $P(X = x)$ antud järgmises tabelis:

x	aa	Aa	AA
$P(X = x)$	0,25	0,5	0,25

NB! Pane tähele: juhuslikke suuruseid tähistatakse enamasti suurte tähtedega $(X, Y, \ldots)$, nende võimalikke väärtuseid väikeste tähtedega $(x, y, \ldots)$.

Kui teame iga juhusliku suuruse X võimaliku väärtuse esinemistõenäosust (ja kui suudame arvutada esinemistõenäosuse ka kõigi võimalike väärtuste kombinatsioonide jaoks), siis öeldakse, et teame juhusliku suuruse X

jaotust. Üks võimalus juhusliku suuruse jaotuse kirjeldamiseks on tõenäosusfunktsioon. Kui teame iga juhusliku suuruse võimaliku väärtuse esinemistõenäosust, siis öeldakse, et teame juhusliku suuruse **tõenäosusfunktsiooni**.

Juhusliku suuruse tõenäosusfunktsiooni saab vahel kirjeldada valemi abil. Uurime näiteks, mitmes apteegis peab narkomaan käima, enne kui leiab apteekri, kes nõustub talle retseptiravimit ilma retseptita müüma. Juhuslikuks suuruseks X on antud juhul apteekide arv, mida narkomaan peab külastama, enne kui ta saab oma tahtmise. Tõenäosus, et narkomaan peab külastama täpselt x apteeki ennem soovitud tulemuseni jõudmist P(X=x), olgu leitav järgmisest valemist:

$$P(X = x) = 0,1 \times 0,9^{x-1}. \tag{3.2}$$

Sellisel juhul ütleme, et juhusliku suuruse X tõenäosusfunktsioon on antud valemiga (3.2) — sest mistahes väärtuse x korral saame leida tõenäosuse, et $P(X = x)$. Soovi korral võime muidugi proovida neid tõenäosuseid ka tabeli kujul esitada, aga see tabel peaks siis olema lõpmatult pikk:

x	1	2	3	4	5	6	7	...
$P(X = x)$	0,1	0,09	0,081	0,0729	0,06561	0,059049	0,053144	...

Vahel on tõenäosusfunktsiooni asemel mugavam või kasulikum kasutada jaotusfunktsiooni $F(x)$. Jaotusfunktsioon kohal x, $F(x)$, on tõenäosus, et juhuslik suurus X ei saa suuremaks kui x, $F(x) := P(X \leq x)$. Jaotusfunktsiooni kutsutakse vahel ka kumulatiivseks jaotusfunktsiooniks.

Järgnevas tabelis on ära toodud nii tõenäosusfunktsiooni ($P(X = x)$) kui ka jaotusfunktsiooni $F(x)$ väärtused ühe ja sama juhusliku suuruse jaoks.

x	1	2	3	4	5	6	7	...
$P(X = x)$	0,1	0,09	0,081	0,0729	0,06561	0,059049	0,053144	...
$F(x) = P(X \leq x)$	0,1	0,19	0,271	0,3439	0,40951	0,468559	0,521703	...

Teades tõenäosusfunktsiooni saab alati leida jaotusfunktsiooni. Teades jaotusfunktsiooni saab samuti leida tõenäosusfunktsiooni (juhul kui juhusliku suuruse jaotust saab üldse tõenäosusfunktsiooni abil kirjeldada — vaata märkust pidevate juhuslike suuruste kohta!).

3.1.3 Jaotuste pere

Millegi poolest sarnased jaotused moodustavad justnagu ühe perekonna. Vahel piisab, kui mainime, millisesse perekonda kõnealune jaotus kuulub, ja

haritud vestluspartner saab juba isegi aru, millega, milliste omadustega jaotusega on tegemist. Igasse jaotuste perekonda kuulub palju jaotuseid. Kui konkreetse inimese leidmiseks peame lisama perekonnanimele inimese eesnime, siis jaotuste puhul peame perekonnanimele lisama kas ühe, või mõnede perede puhul kohe paar numbrit. Jaotuse parameetriteks kutsutakse numbreid (või numbrit), mida teades oskame kõigi perekonda kuuluvate jaotuste seast üles leida just selle ühe ja õige jaotuse.

Kõige tuntumaid jaotuste peresid — nagu näiteks normaaljaotust(e peret), binoomjaotust(e peret) jne — peaks tundma igaüks, kes vähegi tahab statistilist terminooloogiat kasutavast artiklist aru saada või kes ise tahab oma töös kasutada statistilise analüüsi abi.

Bernoulli jaotuste pere

Kui juhuslikul suurusel on kaks võimalikku väärtust, siis kuulub selle juhusliku suuruse jaotus Bernoulli jaotuste sekka. Näiteks kuuluvad Bernoulli perekonda järgmiste juhuslike suuruste jaotused:

Munast kooruva tibupoja soo jaotus

x	0 (=emane)	1 (=isane)
$P(X=x)$	0,492	0,508

Külvatud seemnekese idanemise jaotus

x	0 (=ei idanenud)	1 (=idanes)
$P(X=x)$	0,23	0,77

Kui juhusliku suuruse X jaotus on Bernoulli jaotusega, siis kirjutatakse $X \sim Be(p)$ või $X \sim B(1;p)$. Arv p (tõenäosus, et Bernoulli jaotusega juhuslik suurus omandab väärtuse 1) on Bernoulli jaotuse parameeter (väide $Y \sim Be(0.456)$ määrab üheselt juhusliku suuruse jaotuse).

Binoomjaotuste pere

Oletame, et meie katse "õnnestub" tõenäosusega p. Korraldame n sõltumatut katset. Juhusliku suuruse "õnnestumisega lõppenud katsete arv" (X) jaotus on binoomjaotusega,

$$X \sim B(n,p).$$

Kui juhuslik suurus X on binoomjaotusega $X \sim B(n;p)$, siis tema tõenäosusfunktsioon avaldub kujul

$$P(X=x) = C_n^x p^x (1-p)^{n-x},$$

kus C_n^x näitab, mitmel erineval moel on võimalik n eseme seast valida välja x eset:

$$C_n^x = \frac{n!}{x! \cdot (n-x)!} = \frac{1 \cdot 2 \cdot \ldots \cdot n}{(1 \cdot 2 \cdot \ldots \cdot x) \cdot (1 \cdot 2 \cdot \ldots \cdot (n-x))}.$$

Viimases arvutuses tähistab kirjapilt $n!$ faktoriaali. Tasub meeles pidada, et $0! = 1$.

Näide 3.3 *Pleektatsulaps on emane tõenäosusega 0,6. Üksikul saarel sünnib kaks pleektatsulast. Mitu emast pleektatsulast üksikul saarel sünnib? Milline on üksikul saarel sündivate emaste pleektatsulaste arvu jaotus? Emaste pleektatsulaste arvu jaotus on binoomjaotus* $B(2; 0.6)$*. Leiame selle jaotuse — selleks peame leidma, kui tõenäoliselt sünnib 0, kui tõenäoliselt sünnib 1 ja kui tõenäoliselt sünnib täpselt 2 emast pleektatsut:*

$$\begin{aligned} P(X=0) &= \frac{2!}{0! \cdot (2-0)!} \cdot 0,6^0 \cdot (1-0,6)^{2-0} \\ &= \frac{2}{1 \cdot 2} \cdot 1 \cdot 0,4^2 \\ &= 0,16 \end{aligned}$$

$$\begin{aligned} P(X=1) &= \frac{2!}{1! \cdot (2-1)!} \cdot 0,6^1 \cdot (1-0,6)^{2-1} \\ &= \frac{2}{1 \cdot 1} \cdot 0,6 \cdot 0,4 \\ &= 0,48 \end{aligned}$$

$$\begin{aligned} P(X=2) &= \frac{2!}{2! \cdot (2-2)!} \cdot 0,6^2 \cdot (1-0,6)^{2-2} \\ &= 0,36 \end{aligned}$$

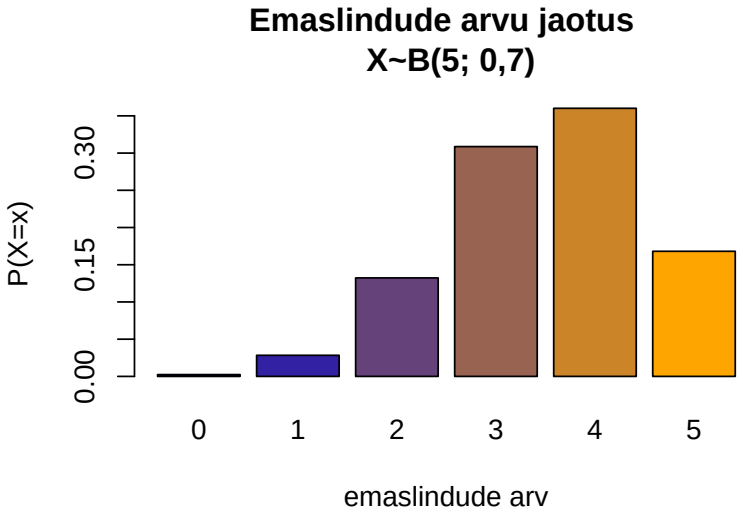
Seega on emaste pleektatsude arvu jaotus, $B(2; 0.6)$*, kirja pandav järgmise tabeli abil:*

x	*0*	*1*	*2*
$P(X=x)$	*0,16*	*0,48*	*0,36*

Näide 3.4 *On varasemast teada, et laborikatse õnnestumise tõenäosus on 0,7. Tudengil on aega ja vahendeid 5 katse tegemiseks. Juhusliku suuruse* X *— õnnestunud katsete arvu — jaotus on binoomjaotus,* $X \sim B(5; 0,7)$*:*

x	*0*	*1*	*2*	*3*	*4*	*5*
$P(X=x)$	*0,00243*	*0,02835*	*0,1323*	*0,3087*	*0,36015*	*0,16807*

Antud jaotust iseloomustab ka joonis 3.2.



Joonis 3.2: Binoomjaotusega juhusliku suuruse $X \sim B(5; 0.7)$ jaotus

Veel kuulsaid diskreetsete tunnuste jaotuseid

- Poissoni jaotus. Poissoni jaotusega on näiteks ühe päeva jooksul aset leidvate südameatakkide arv Tartu linnas, raku jagunemisel tekkivate geenimutatsioonide arv jne. Seda, et tunnus X on Poissoni jaotusega, tähistatakse $X \sim P(\lambda)$, kus λ on keskmine südameatakkide arv ühes päevas või keskmine mutatsioonide arv raku jagunemisel;
- Geomeetriline jaotus – kui katse õnnestumise tõenäosus on p, siis katsete arv kuni esimese õnnestumiseni on geomeetrilise jaotusega juhuslik suurus; näide, kus juhuslikuks suuruseks oli apteekide arv, mida narkomaan pidi külastama enne soovitud tulemuseni jõudmist oli näide geomeetrilisest jaotusest parameetriga 0,1.
- . . . .

Pideva tunnuse jaotus

Pideva tunnuse korral saab rääkida tunnuse jaotusfunktsioonist $F(x) = P(X \leq x)$, ehk tõenäosusest, et juhuslikult valitud objektil uuritava tunnuse väärtus on samasuur või väiksem arvust x. Tõenäosusfunktsiooni aga kasutada ei saa. Probleem seisneb nimelt selles, et pideva tunnuse mistahes

väärtuse esinemistõenäosus on null (millise tõenäosusega on teile tänaval vastutuleva inimese pikkus täpselt 164,59210001235295867219002....cm?). Üksikväärtuste tõenäosuste asemel vaadeldakse pidevate tunnuste korral, kui suur on tõenäosus, et juhuslik suuruse väärtus satub mingisse lõiku a-st b-ni. Sellist funktsiooni $f(x)$, mille graafiku alune pindala lõigus a-st b-ni on alati võrdne tõenäosusega, et juhuslik suurus omandab väärtuse selles vahemikus (mistahes a ja b valiku korral), nimetatakse tihedusfunktsiooniks. Matemaatiliselt kirjapandult näeb sama nõue välja nii:

$$P(a < X \leq b) = \int_a^b f(x)dx$$

Vaata ka joonist 3.3, kus on esitatud ühe juhusliku suuruse jaotus- ja tijedusfunktsioon.

Teades tihedusfunktsiooni, on võimalik leida jaotusfunktsiooni:

$$F(y) = P(X \leq y) = \int_\infty^y f(x)dx,$$

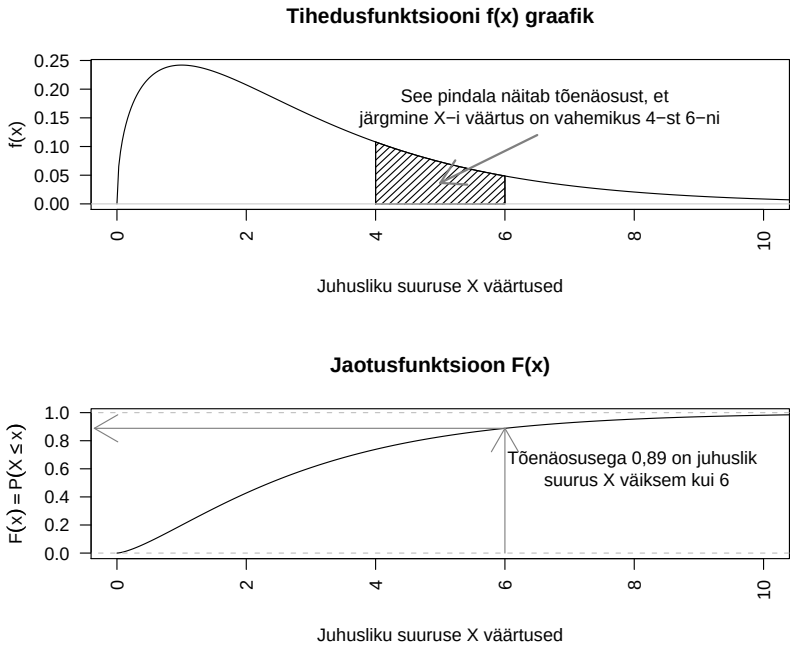
ja vastupidi — jaotusfunktsiooni teades saame leida tihedusfunktsiooni:

$$f(x) = \frac{\partial F(x)}{\partial x}.$$

Märkus: Vahel kasutatakse ka nn elulemusfunktsiooni (elukestvusfunktsiooni) $S(x) := 1 - F(x) = P(X > x)$, mis on lihtsalt jaotusfunktsiooni teisend.

Seos tihedusfunktsiooni ja histogrammi vahel

Pideva juhusliku suuruse histogrammi oli võimalik konstrueerida, jagades tunnuse väärtused intervallidesse. Kui mitu korda juhuslik suurus sattus mingisse antud intervalli, seda kõrgem tulp tuli antud intervalli kohale joonistada. Pideva tunnuse puhul saab intervalle moodustada mitut moodi ja sõltuvalt intervallide valikust võib ka tunnuse histogramm märgatavalt muutuda. Suure valimi korral võib ka küllalt kitsastesse intervallidesse sattuda palju vaatluseid ja sellisel juhul säilitab pideva tunnuse histogramm oma üldkuju sõltumata täpselt intervallide valikust (eeldades siiski, et valitud intervallid on samal histogrammil kõik sama laiad). Kui valim on väga suur, ja on võimalik intervallid teha üsna kitsad, siis hakkavad pisikeste tulpade otsad moodustama kõverjoont, mis osutub kujult väga sarnaseks populatsiooni tihedusfunktsioonile.



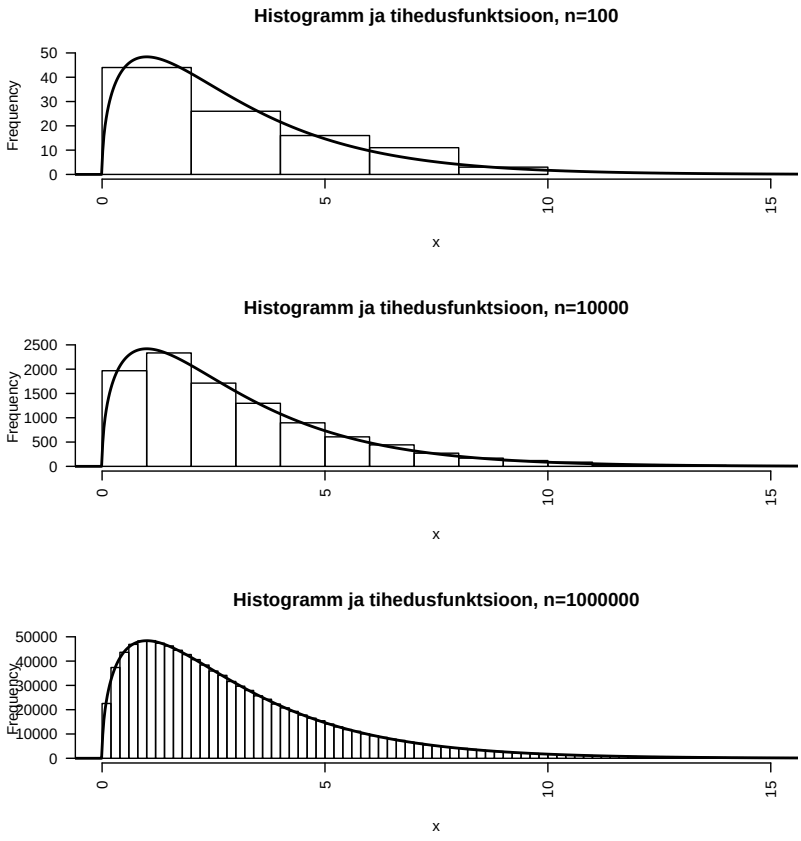
Joonis 3.3: Ühe pideva juhusliku suuruse tihedus- ja jaotusfunktsioon

Näide 3.5 *Joonistele 3.4 on kantud pideva musta joone abil ühe juhusliku suuruse tihedusfunktsioon (korrutatud läbi konstandiga— et ta sattuks samasse skaalasse histogrammiga) ja kolm erinevat histogrammi, igal korral erinevat intervallilaiust kasutades. Võime märgata, et suure valimi (ja väikeste intervallide) korral on histogrammi kuju üsna sarnane tihedusfunktsioonile.*

Normaaljaotus

Üks sagedamini ette tulev jaotus(te pere) elusloodeses. Kui uuritavat tunnust mõjutavad paljud erinevad tegurid, millest ühegi mõju pole omaette võttes märkimisväärne, siis on uuritava tunnuse jaotus sageli lähedane normaaljaotusele. Näiteks kipuvad olema normaaljaotusega paljude geenide poolt määratavad näitajad: pikkus, kaal, lehmade piimaand, ... Matemaatiliselt öeldult: paljude juhuslike suuruste summa jaotuseks on (ligilähedaslt) normaaljaotus. Enamasti leitakse ka, et mõõtmisvigade jaotus kipub olema normaaljaotus.

Kui uuritava tunnuse jaotus on normaaljaotusega, siis tema tihedusfunkt-



Joonis 3.4: Tihedusfunktsioon ja histogram

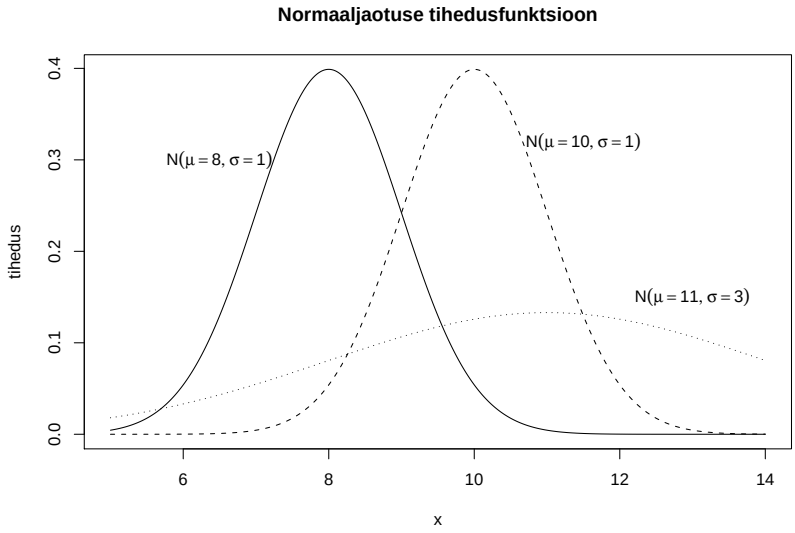
sioon on esitatav järgmisel kujul:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

Normaaljaotusel on kaks jaotusparameetrit — μ on uuritava tunnuse keskväärtus ($\approx$ uuritava tunnuse kõigi väärtuste keskmine) ja σ^2 on uuritava tunnuse dispersioon. Jaotuse täpseks määramiseks piisab seega, kui me saame öelda: uuritav tunnus on normaaljaotusega, sellise-ja-sellise keskväärtuse ja dispersiooniga.

Kolme normaaljaotusega juhusliku suuruse tihedusfunktsioonid on ära toodud joonisel 3.5.

Normaaljaotust, mille keskväärtus on 0 ($\mu = 0$) ja standardhälve on 1



Joonis 3.5: Normaaljaotuse tihedusfunktsioon

($\sigma = 1$), kutsutakse standardseks normaaljaotuseks. Standardse normaaljaotuse jaotusfunktsiooni tähistatakse sageli sümboliga $\Phi(x)$.

Normaaljaotust esineb palju ka seetõttu, et "normaalsust" on raske hävitada. Olgu meil algse juhusliku suuruse (ehk tunnuse) jaotuseks normaaljaotus, $X \sim N(\mu, \sigma^2)$. Siis me teisendame oma juhuslikku suurust kuidagi (näiteks logaritmime, juurime vms) ja vaatame uut juhuslikku suurust $Y = g(X)$. Üllataval kombel selgub, et ka teisendatud väärtuste jaotuseks on ligikaudu normaaljaotus (juhul kui $g'(\mu) \neq 0$),

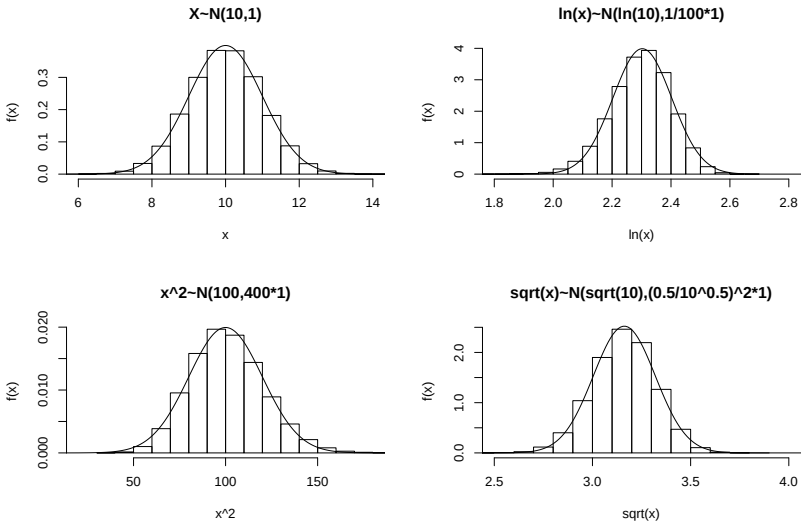
$$Y \sim N(g(X); \{g'(\mu)\}^2\sigma^2).$$

Seda fenomeeni iseloomustab ka joonis 3.6.

Kui X on normaaljaotusega, keskväärtusega μ ja dispersiooniga σ^2, siis pole kerge leida tõenäosust tihedusfunktsiooni integreerimise teel. Selle asemel kasutatakse standardse normaaljaotuse jaotusfunktsiooni tabeleid või arvutitarkvara.

Kui $X \sim N(\mu, \sigma^2)$, siis $\frac{X-\mu}{\sigma} \sim N(0, 1)$. Seega

$$\begin{aligned} P(a < X \leq b) &= P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \end{aligned}$$



Joonis 3.6: Normaaljaotuse tihedusfunktsioon

Kui valim on normaaljaotusega, siis ligikaudu 68,3% vaatlustest jäävad vahemikku $\mu \pm \sigma$, 95,5% väärtustest jäävad vahemikku $\mu \pm 2\sigma$ ja 99,7% vahemikku $\mu \pm 3\sigma$.

Normaaljaotus on statistikas erilise tähtsusega, sest:

1. Paljud valimid on ligikaudu normaaljaotusega. Näiteks juhul, kui uuritavat tunnust mõjutavad paljud erinevad tegurid, millest ühegi mõju omaette pole tugev, siis on uuritava tunnuse jaotus lähedane normaaljaotusele. Seega tunnused, mis on määratud väga paljude geenide poolt, on enamasti normaaljaotusele väga lähedase jaotusega (pikkus, kaal, lehmade piimaand,...)

2. Väga paljud statistilise analüüsi meetodid eeldavad normaaljaotusega valimit.