Peatükk 4

Valim ja populatsioon

Õpime tükikeste põhjal ette kujutama tervikut
ehk
Valikust, valimist ja esindavast valimist

Oma uurimistööd tehes uurime enamasti läbi vaid killukese meid huvitavatest objektidest. Soovides teada, kui levinud on ebaseaduslikud metsaraied, jõuame ehk üle vaadata vaid paarsada metsatukka - aga nende paarisaja koha põhjal tahaksime kangesti öelda midagi ebaseadusliku metsaraie kohta Eestis tervikuna. Või uurides kartuli viirusnakkuse levikut võime võtta ehk proove sajalt põllult ning määrata, kas elujõulist viirust esineb meie proovides või mitte - aga suurem huvi oleks nähtavasti ikka öelda, kas antud viirushaigus on käesoleval aastal Eestis laialt levinud või mitte (ja seda väites tahame, et meie väide kehtiks ikka kõigi põldude kohta). Ning kui püüame kinni 10 jänest ja kirjeldame neid, siis on meie lootuseks ikka see, et sedaviisi saame kirjeldada jäneseid tervikuna, jänest kui liiki. *Populatsioon* on kõigi objektide, isendite, esemete, nähtuste või seisundite kogum, mille kohta soovitakse järeldusi teha. Sageli kasutatakse populatsiooni asemel ka *üldkogumi* mõistet. Populatsiooni defineerides piiritletakse ära uuritav objekt (ruumis, ajas, katsetingimuste kaudu,...).

Näiteid populatsioonidest: Eesti talud aastal 2007; tõve X käes vaevlevad lehmad (nii minevikus, praegu kui ka tulevikus); kõik antud mündiga teha võidavad kulli/kirja viskamised, ...

Kui ei suudeta täpselt kirjeldada populatsiooni, kelle kohta midagi tahetakse väita, siis on esitatud väide ka suuresti kasutu. Näiteks väide: väetamine väetisega XYZ tõstab nisu saagikust kaks korda on vaid eksitava tähtsusega, kui ei teata, millise populatsiooni kohta antud väide kehtib (nisu kasvatamisel troopilisel Borneo saarel vihmaperioodil piirkondades, kus

uugu-mangod võivad vabalt ringi liikuda ja põldudelt vilja süüa — nimelt muudab vastav väetis nisu uugu-mangodele mürgiseks ja seetõttu jääb rohkem saaki inimestele).

Üldkogumi neid objekte, mida on vaadeldud või uurimiseks välja valitud, kutsutakse *valimiks*.

Populatsiooni kohta järelduste tegemiseks pole kõik valimid ühtmoodi head. Soovides näiteks uurida, milline on Eesti talude majanduslik seisund, siis on vähe kasu, kui meie kasutada on andmed kümne hiljuti pankrotistunud talu kohta. Hoopis parem oleks valim, kus virelevaid ja õitsvaid talusid oleks ligikaudu samas proportsioonis kui populatsioonis tervikuna. Valimit, kus uuritava tunnuse jaotus on enam-vähem samasugune kui populatsioonis, nimetatakse *esindavaks*. Hea valimi saamiseks on väga tähtis valimi moodustamise eeskiri - valikueeskiri ehk -disain.

Parim juht on loomulikult siis, kui valim ja populatsioon kattuvad. Sellisel juhul on valimi esindav ja taolisel valimil teostatud uuringut nimetatakse *kõikseks uuringuks*. Paraku tuleb kõikseid uuringuid elus harva ette. Näiteks välistab kõikne uuring oma loomult igasugused teaduslikel alustel tehtavad tulevikuprognoosid (kui populatsioon haaraks ka tulevikus eksisteerivaid objekte/sündmuseid – näiteks järgmisel aastal sündivaid jäneseid – siis peaks kõikse uuringu korral olema meie andmestikus andmed ka järgmisel aastal sündivate jäneste kohta).

Üks kindlamaid viise esindava valimi saamiseks on juhuslik valik. Nimelt valides objekte valimisse täiesti juhuslikult ja moodustades sellisel põhimõttel piisavalt suure valimi, saame esindava valimi. Korrektselt moodustatud juhusliku valimi korral peab igal objektil (isendil) olema võrdne võimalus sattuda valimisse (pärapõrgus asuval talul peab olema samasuur tõenäosus sattuda valimisse kui uuringut tegeva teadlase naabril). Samuti ei tohiks tõenäosus sattuda valimisse sõltuda sellest, kes juba valimisse on sattunud (me ei tohiks võtta valimisse sattunud talu ümbert veel talusid oma valimisse, igal talul peaks olema teistest sõltumatu võimalus valimisse sattuda ja naabrid ei tohiks teineteist aidata). Kui mainitud eeldused on täidetud, siis sarnaneb valimi jaotus populatsiooni jaotusele (mida suurem valim, seda sarnasem). Iseloomustamaks toodud väidet, on ära toodud tabel 4.1, kus on kirjas populatsiooni jaotus (mida me üldjuhul ei tea) ja nelja erineva suurusega juhusliku valimi jaotused.

Tabel 4.1: Populatsiooni ja valimi jaotus

Uuritava tunnuse väärtused	Populatsiooni jaotus	Valimi jaotus			
		n=20	n=40	n=100	n=500
Hall	60%	45%	62,5%	62,0%	60,0%
Must	35%	55%	30,0%	36,0%	34,8%
Valge	5%	0%	7,5%	2,0%	5,2%

Peatükk 5

Populatsiooni parameetrite hindamine. Hinnangu viga

Enamasti pakuvad uurijale huvi populatsiooni iseloomustavad näitajad, mitte aga valimit iseloomustavad statistikud. Näiteks võib meid huvitada, kui kõrge on sordi XYZ idanemisprotsent, aga see, kui mitu sordi XYZ tera läheb idanema meie valimis, huvitab meid vaid sedavõrd, kuivõrd ta aitab vastata üldisemale, populatsiooni puudutavale küsimusele.

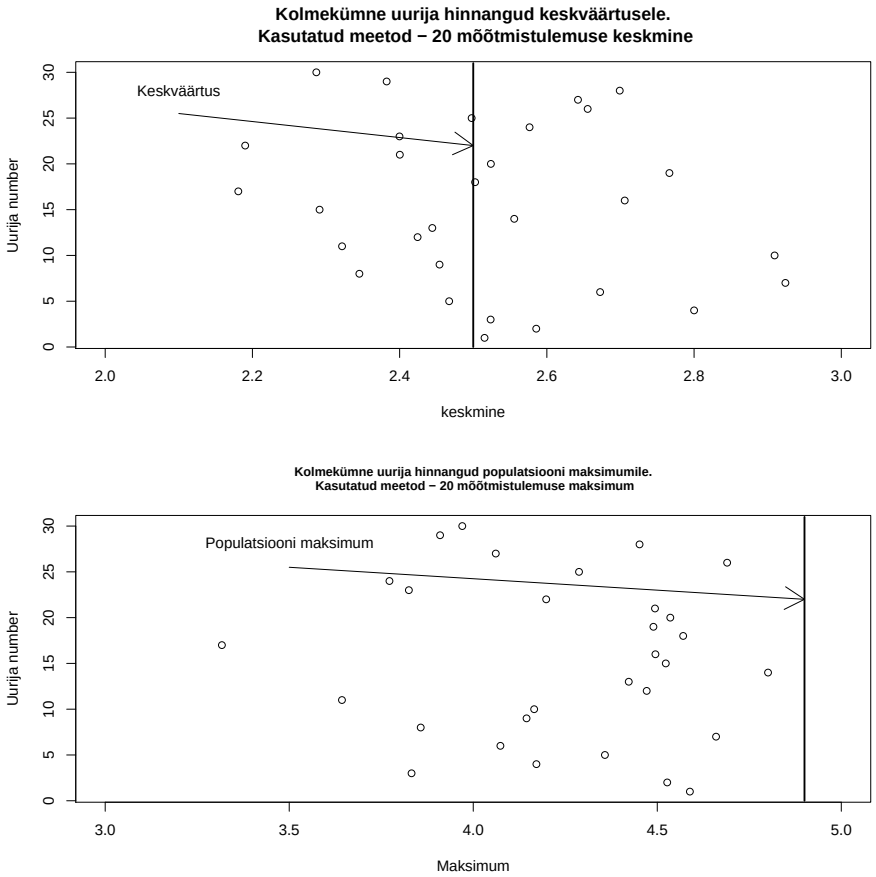
Populatsiooni parameetrite hindamiseks on mitmeid võimalusi. Tegelikku idanemisprotsenti võime hinnata kasutades kohvipaksu, valimi põhjal midagi arvutades või kasutades veel mõnda muud lähenemisviisi (näiteks alati pakkudes välja väärtust 3).

Saades hinnangud kolmel erineval viisil (kolm numbrit) võib olla väga raske öelda, milline neist kolmest numbrist on parim (näiteks lähedaseim õigele väärtusele). Täiesti võimalik, et seekord andis täpseima hinnangu ekspert, kes igale võimalikule küsimusele pakub vastuseks numbrit 3. Siiski on võimalik katsetada erinevaid hindamismeetodeid olukordades, kus me õiget vastust teame, ja vaadata, kui hästi erinevad hindamismeetodid suudavad õiget vastust ära arvata. Hiljem võime siis ehk meelsamini usaldada sellist metoodikat, mis võrreldes teistega tuntud olukordades paremini on töödanud. Seega valime hindamismetoodika selle järgi, millised on meetodi kui sellise omadused, ja mitte selle järgi, milline meetod meie valimi põhjal annab täpseima vastuse (seda me lihtsalt ei tea).

Millised peaksid olema hea hindamismeetodika omadused? Kaks olulist omadust on nihketus (nihkega hinnang võib pakkuda hinnanguid, mis kipuvad õigest väärtusest näiteks alatihti suuremad olema, nihketa hinnang aga ei tee süstemaatilist ehk tahtlikku viga üheski suunas) ja omadus anda

olemasoleva informatsiooni põhjal kõige täpsemaid hinnanguid (efektiivsus).

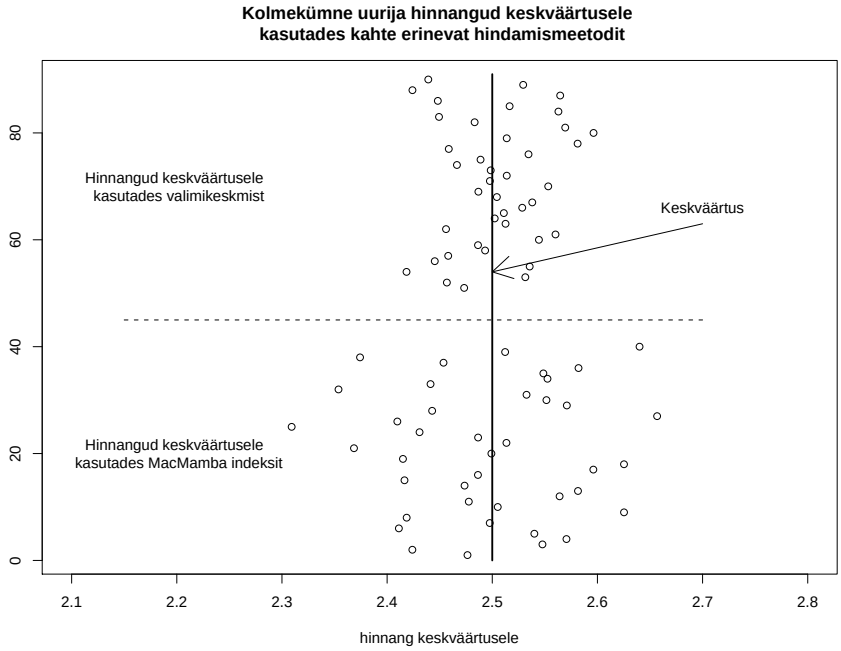
Joonis 5.1: Näide nihketa ja nihkega hinnangust



Populatsiooni väärtustest rääkimisest tuleb õppida vahet tegema tegelikul, aga meile teadmata väärtusel ja ühel või mitmel selle populatsiooni väärtuse hinnangul. Sagedamini kasutavate näitajate jaoks on isegi välja mõeldud erinevad tähistused. Näiteks populatsiooni keskmise ehk keskväärtuse tähistamiseks kasutatakse sageli järgmiseid sümboleid: $\mathrm{E}X$, μ; valimi keskväärtust tähistatakse aga sümboliga $\bar{x}$. Tabelis 5.1 on ära toodud populatsiooni parameetrite enamlevinud tähistused ja nende hindamiseks kasutatavate valimi näitajate nimed ja tähistused.

Kui huvipakkuva väärtuse hinnanguks on üks konkreetne arv, näiteks kui hindame keskväärtust kasutades valimi keskmist, siis räägitakse, et tegemist

Joonis 5.2: Näide täpsemast ja vähemtäpsemast meetodist



Tabel 5.1: Populatsiooni parameetreid ja nende hinnanguid

Populatsiooni parameeter	hinnang valimi põhjal
keskväärtus $\mathrm{E}X$, μ	keskmine $\bar{x}$
populatsiooni dispersioon $\mathrm{D}X$, σ^2	valimi dispersioon s^2
populatsiooni standardhälve σ	valimi standardhälve s
populatsiooni mediaan $\mathrm{med}(X)$	valimi mediaan $\hat{med}(X)$
populatsiooni α-kvartiil q_α	valimi α-kvartiil $\hat{q}_\alpha$

on punkthinnanguga. Märkimaks, et tegemist on hinnanguga, kirjutatakse sageli arvkarakteristikut iseloomustava sümboli kohale laineke, katuseke,

tärn vms.

5.1 Punkthinnangu viga

Kuna iga teadlane kasutab populatsiooni kirjeldamiseks erinevat juhuslikku valimit, siis erinevate uurijate poolt antud populatsiooni kirjeldused (hinnangud) paraku ei kattu teineteisega (ning erinevad ka populatsiooni tegelikest parameetritest). Illustreerimaks seda väidet toome järgmise näite.

Näide 5.1 *Tuntakse huvi rannateo tööstusliku kasvatamise võimaluste vastu Eestis. Üks oluline parameeter, mida kasvatustiikide planeerimisel teadma peab, on see, mitu järglast rannatigu kasvatustiigis Eesti oludes keskmiselt ilmale toob. Saamaks informatsiooni järglaste arvu keskväärtuse kohta, luges vapper doktorant Juhan Kajakas üle 100 rannateo järglased. Oma valimi põhjal leidis Juhan Kajakas, et rannateol on keskmiselt 28,8 järglast.*

Samas tunnevad Eestis rannateo kasvatuse avamise vastu huvi ka hiinlased. Nii saabus siia Hung-Hang, väärikas Pekingist pärit teadlane ja luges samuti üle 100 rannateo järglased ning sai oma valimi keskmiseks 30,6. Peagi olid kohal ka teadlased teistest piirkondadest ning igaüks otsis sama metoodika alusel vastust samale küsimusele - kui palju järglaseid annab rannatigu keskmiselt Eesti oludes. Igaüks neist sai vastuseks veidi erineva numbri. Saadud hinnangud on esitatud joonisel 5.1.

Teades kõigi nende uuringute tulemusi, on lihtne iseloomustada uuringu metoodika täpsust - kasutades näiteks uuringukeskmiste dispersiooni või standardhälvet, saame kirjeldada, kui kaugele võivad erinevad sama metoodikat kasutavate uuringute tulemused teineteisest tulla. Paraku pole meil uuringut teostades enamasti võimalik kasutada teiste sarnaste uuringute tulemusi (kui küsimust oleks juba uuritud, oleks rahastajate huvi antud küsimuse vastu juba märksa väiksem...). Üldjuhul pole ühe juhusliku suuruse väärtuse põhjal võimalik hinnata tema dispersiooni (Miks?). Uurides aga hoolikalt dispersiooni omadusi, selgub, et uuringukeskmise dispersiooni leidmine on võimalik ka siis, kui tehtud on kõigest üksainus uuring.

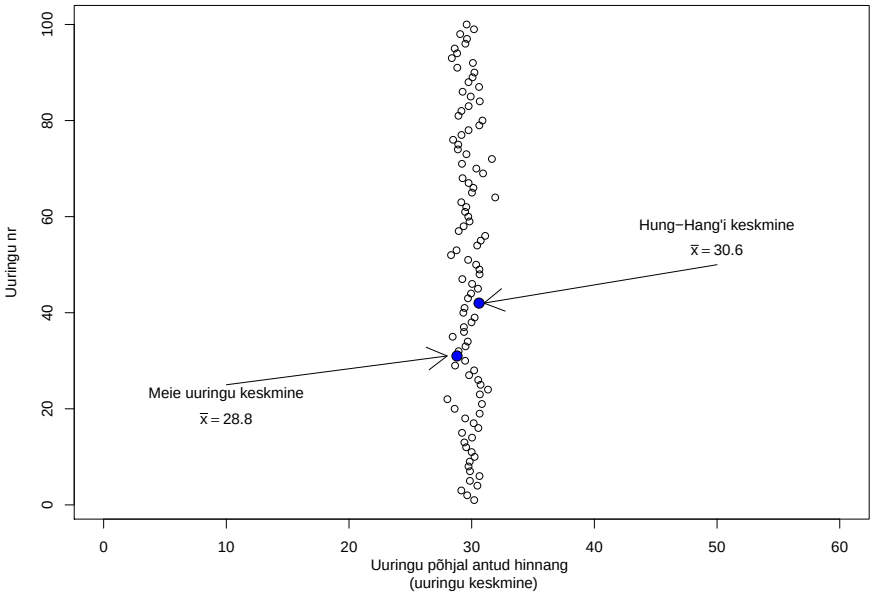
5.1.1 Dispersiooni omadusi

Populatsiooni dispersioon defineeritakse kui uuritava tunnuse väärtuste keskmine ruutkaugus keskväärtusest:

$$DX := E(X - EX)^2 = E(X^2) - (EX)^2.$$

Dispersiooni omadusi:

Joonis 5.3: Saja sama metoodikaga tehtud uuringu tulemused



1. Juhusliku suuruse X ja konstandi c korral $D(cX) = c^2 DX$;

2. Juhusliku suuruse X ja konstandi c korral $D(X + c) = DX$;

3. Kui juhuslikud suurused X ja Y on sõltumatud, siis $D(X + Y) = DX + DY$.

4. Kui uuritava tunnuse X dispersioon populatsioonis on σ^2, siis valimi keskmise dispersioon on σ^2/n, kus n tähistab valimi suurust.

Viimase väite ka tõestame:

$$\begin{aligned} D(\bar{X}) &= D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &\overset{(1)}{=} \frac{1}{n^2} D\left(\sum_{i=1}^{n} X_i\right) \\ &\overset{(3)}{=} \frac{1}{n^2}\sum_{i=1}^{n} D\left(X_i\right) \\ &= \frac{1}{n^2} n\sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Kasutades dispersiooni omadust 4 saame (korrektse juhusliku valimi korral) leida hinnangu valimi keskmise dispersioonile, ilma, et peaksime teadma teiste uurijate tulemusi. Nimelt oskame hinnata populatsiooni dispersiooni σ^2, kasutades valimi dispersiooni s^2. Seega hinnates keskväärtust kasutades valimi keskmist, oskame kirjeldada oma hinnangu täpsust - valimi keskmise dispersiooni hinnang on s^2/n.

Lisaks tasub märgata, et suure valimi korral on valimi keskmise jaotuseks normaaljaotus. Nimelt paljude enam-vähem võrdse suurusega juhuslike suuruste summa jaotuseks on normaaljaotus. Seetõttu saame väita, et suure valimi korral on valimi keskmise jaotuseks normaaljaotus:

$$\bar{X} \sim N(\mu; \sigma^2/n).$$

5.2 Standardviga

Parameetri hinnangu standardhälvet nimetatakse standardveaks — inglise keeles *standard error*, lühendina kasutatakse sageli tähekombinatsioone *se* või *s.e.*:

$$s.e. = \sqrt{\hat{D}(\bar{X})} = \sqrt{s^2/n} = s/\sqrt{n}.$$

Sageli esitatakse teaduskirjanduses hinnatud parameeter (näiteks valimi keskmine) koos standardveaga, sageli kujul *keskmine* $\pm$ *standardviga* või *keskmine*(*standardviga*). Näiteks: sort A saagikus oli $12,3 \pm 0,7$ tonni/ha.

Hoiatus! Sama kirjapilti kasutades lisatakse vahel keskmise taha hoopistükkis uuritava tunnuse standardhälve. Sestap tuleks ise artiklit kirjutades

kuskil ära märkida, mida käesolevas artiklis keskmise taha kirjutatud arvud tähendavad – kas standardviga või standardhälvet.

Kumba numbrit, kas standardviga või standardhälvet peaksin mina oma artiklis keskmise taha kirjutama? Vastus sõltub mõnevõrra sellest, mida tahetakse artikli lugejale öelda. Kui soovitakse enam edasi anda algse tunnuse väärtuste hajuvust (kui mina külvaksin oma põllule seemet sordist A, siis kuivõrd erineva saagi ma keskmisest võin saada), siis oleks soovitav kasutada standardhälvet. Kui aga põhitähelepanu on keskväärtuste võrdlemisel (kas sort A saagikuse keskväärtus on ikka parem sort B või sort C saagikuse keskväärtusest), on soovitavam lisada keskmiste taha hinnangu täpsust kirjeldav standardviga.

Peatükk 6

Prognoosiintervall ja Usaldusintervall

6.1 Prognoosiintervall

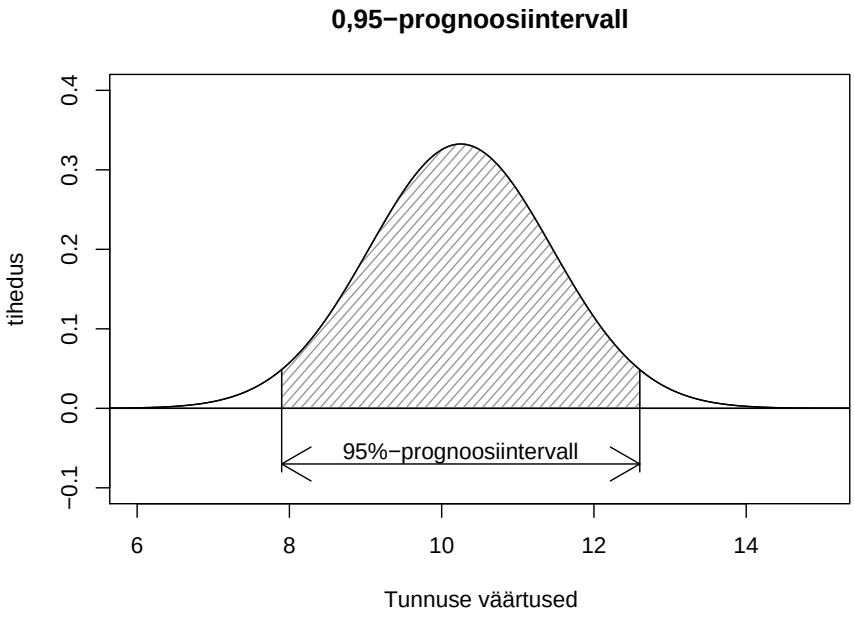
Unustame hetkeks populatsiooni parameetrite hindamise ja pöördume tagasi üksikvaatluste juurde.

On raske ennustada, milline on huvipakkuva tunnuse väärtus järgmisel juhuslikult populatsioonist valitud isendil. Samas on võimalik leida vahemik, millese järgmine vaatlus satub suure tõenäosusega. Näiteks, kui on teada uuritava tunnuse jaotus populatsioonis (teame sigade kaalu tihedusfunktsiooni), siis võib valida vahemiku, kuhu uuritava tunnuse väärtus (ühe juhuslikult valitud sea kaal) sattub mingi suure etteantud tõenäosusega. Alltoodud graafikul 6.1 on konstrueeritud selline prognoosintervall, kuhu järgmine üksikvaatlus sattub tõenäosusega 0,95 (95%-prognoosiintervall).

Kuidas prognoosipiire reaalselt saaks leida? Üks võimalus on muidugi kasutada sobivalt valitud kvantiile — 0,025-kvantiili ja 0,975-kvantiili vahele jääb uuritava tunnuse väärtus tõenäosusega $0,975 - 0,025 = 0,95$, seega on nende kahe kvantiili poolt määratud vahemik 0,95-prognoosiintervall. Üldjuhul läheks meil $(1-\alpha)$-prognoosiintervalli leidmiseks vaja teada $\alpha/2$ ja $(1-\alpha/2)$-kvantiile — nende kahe kvantiili vahele jääb juhusliku suuruse väärtus täpselt tõenäosusega $(1-\alpha)$.

Vaatame juhtu, mil uuritava tunnuse jaotuseks on standardne normaaljaotus (selline normaaljaotus, mille puhul $\sigma^2 = 1$ ja $\mu = 0$). Sellisel juhul võime tabelist 6.1 välja lugeda soovitud (standardse normaaljaotuse) kvantiilid ning leida soovitud prognoosiintervalli. Märkus: standardse normaaljaotuse α-kvantiili tähistatakse traditsiooniliselt sümboliga z_α.

Joonis 6.1: 0,95-prognoosiintervall



Tabel 6.1: Standardse normaaljaotse kvantiilid z_α

α	0,005	0,025	0,05	0,5	0,95	0,975	0,995
z_α	-2,58	-1,96	-1,64	0	1,64	1,96	2,58

Seega standardse normaaljaotusega juhusliku suuruse korral oleks 0,95-prognoosiintervall (-1,96...1,96); 0,9-prognoosiintervall (-1,64...1,64) ja 0,99-prognoosiintervall (-2,58...2,58).

Uuritavaid tunnuseid, mille jaotuseks oleks täpselt standardne normaaljaotus, esineb tavaelus haruharva. Küll aga esineb normaaljaotusega juhuslikke suuruseid, ja mitte harva.

Kuidas leida prognoosiintervalli normaaljaotusega juhuslikule suurusele X, mille keskväärtus $EX = \mu$ ja dispersioon on $DX = \sigma^2$, ehk teisisõnu, mille jaotuseks on $X \sim N(\mu, \sigma^2)$?

Esmalt märkame, et teisendatud juhusliku suuruse $X_{uus} := \frac{X-\mu}{\sigma}$ keskväärtuseks on 0 ja dispersiooniks 1:

$$EX_{uus} = \frac{1}{\sigma}(EX - \mu) = \frac{1}{\sigma}(\mu - \mu) = 0;$$
$$DX_{uus} = D\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}D(X-\mu) = \frac{1}{\sigma^2}\sigma^2 = 1.$$

Kuna normaaljaotusega juhuslikule suurusele konstandi juurdeliitmisel või konstandiga korrutamisel saame ikka normaaljaotusega juhusliku suuruse, siis järelikult on uue, teisendatud juhusliku suuruse X_{uus} jaotuseks standardne normaaljaotus ja 95% tema väärtustest jääb vahemikku -1,96..1,96:

$$\begin{aligned} P(-1,96 \le X_{uus} \le 1,96) &= 0,95 \\ P(-1,96 \le \frac{X-\mu}{\sigma} \le 1,96) &= 0,95 \\ P(-1,96\sigma \le X-\mu \le 1,96\sigma) &= 0,95 \\ P(\mu - 1,96\sigma \le X \le \mu + 1,96\sigma) &= 0,95 \end{aligned}$$

Järeldus: Kui juhusliku suuruse X jaotuseks on normaaljaotus, $X \sim N(\mu, \sigma^2)$, siis tema 0,95-prognoosiintervall on leitav järgmise valemiga:

$$(\mu - 1,96\sigma \ldots \mu + 1,96\sigma).$$

Loomulikult võime sama arutluskäiku korrata, otsides teisendatud juhuslikule suurusele X_{uus} mingit muud, näiteks $1-\alpha$-prognoosiintervalli. Tulemuseks saame:

$$\left(\mu + z_{\alpha/2}\sigma \ldots \mu + z_{1-\alpha/2}\sigma\right). \tag{6.1}$$

Seega saame standardse normaaljaotuse kvantiile teades leida suvalist prognoosiintervalli suvalise normaaljaotusega juhuslikule suurusele — senikaua kuni teame meid huvitava juhusliku suuruse keskväärtust ja standardhälvet.

Märkus: Mis saab siis, kui me ei tea keskväärtust ja populatsiooni standardhälvet σ? Suure valimi korral võime muidugi teha näo ja väita, et meie valimi keskmine ja standardhälve on väga-väga täpsed hinnangud populatsiooni parameetritele, peaaegu võrdsed nendega, ja seega võime kasutada ka valemit 6.1, asendades vaid μ valimi keskmisega $\overline{x}$ ja σ valimi standardhälbega s. Väiksemates valimites võib erinevus valimi põhjal saadud hinnangute $(\overline{x}, s)$ ja populatsiooni väärtuste (μ, σ) vahel olla siiski märkimisväärne. Sellisel juhul tuleks $(1-\alpha)$-prognoosiintervall normaaljaotusega juhuslikule suurusele leida kasutades valemit:

$$\left(\overline{X} + t_{\alpha/2;n-1}S\sqrt{1+\frac{1}{n}} \ldots \overline{X} + t_{1-\alpha/2;n-1}S\sqrt{1+\frac{1}{n}}\right). \tag{6.2}$$

Kus $t_{\alpha/2;n-1}$ ja $t_{1-\alpha/2;n-1}$ on vastavalt t-jaotuse $\alpha/2$ ja $1-\alpha/2$-kvantiilid. Tasub ehk ära märkida, et kui prognoosiintervall 6.1 on tõlgendatav kui vahemik, kuhu vahele jääb $(1-\alpha)\cdot 100\%$ vaatlustest, siis valemiga 6.2 kirjeldatud prognoosivahemik näitab küll vahemikku, kuhu sattub järgmine vaatlus tõenäosusega $(1-\alpha)$, aga kuhu ei pruugi jääda $(1-\alpha)$ osa kõigist tulevastest vaatlustest (ligikaudu on väide siiski õige, aga mitte päris täpselt)!

6.2 Usaldusintervall

Ka punkthinnangud on juhuslikud suurused — sest valim on juhuslik. Iga uurija, kes üritab vastata samale (populatsiooni puudutavale) küsimusele, saab veidi teistest erineva vastuse. Kui lähemalt uurida selle juhusliku suuruse - punkthinnangu - jaotust, siis selgub üllatavalt sageli, et tegemist on normaaljaotusega. Näiteks on vähegi suurema valimi korral (kümmekond või enam vaatlust) valimi keskmise jaotuseks normaaljaotus:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Seega on teisendatud juhusliku suuruse $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ jaotuseks standardne normaaljaotus. Standardse normaaljaotuse korral aga oskame leida vahemikku, kuhu vahele standardse normaaljaotusega juhusliku suuruse väärtus peab sattuma tõenäosusega 0,95:

$$\begin{aligned} P\left(-1,96 \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le 1,96\right) &= 0,95 \\ P\left(-1,96\frac{\sigma}{\sqrt{n}} \le \bar{X}-\mu \le 1,96\frac{\sigma}{\sqrt{n}}\right) &= 0,95 \\ P\left(\bar{X}-1,96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}+1,96\frac{\sigma}{\sqrt{n}}\right) &= 0,95 \end{aligned}$$

Seega 95% juhuslike valimite korral jääb populatsiooni tegelik keskväärtus vahemikku

$$\left(\bar{X}-1,96\frac{\sigma}{\sqrt{n}} \ldots \bar{X}+1,96\frac{\sigma}{\sqrt{n}}\right).$$

Antud juhul tasub tähele panna järgmist: juhuslik pole mitte populatsiooni tegelik keskväärtus μ, vaid leitud vahemik: iga uurija võib teisest saada veidi erineva vahemiku. Ühte konkreetset valimit (ja valimi keskmist) kasutades leitud vahemik kas sisaldab või ei sisalda populatsiooni keskväärtust.

Tõenäosusest konkreetse, väljaarvutatud intervalli kontekstis enam rääkida ei saa. Küll aga saab rääkida, et kasutasime arvutusmetoodikat, mis 95%-l juhtudest annab populatsiooni tegelikku keskväärtust sisaldava vahemiku ja seega võime 95%-se kindlusega (*confidence*) väita, et tegelik keskväärtus asub antud vahemikus. Vastavat vahemikku kutsutakse 95%-usaldusintervalliks või 95%-usaldusvahemikuks.

Hakates oma uuringu tarvis taolist 95%-st vahemikku leidma, põrkume aga raskuse otsa — me ei tea ju populatsiooni standardhälvet σ. Esimene mõte, mis pähe võiks tulla, oleks järgmine — asendame σ tema hinnanguga, valimi standardhälbe s-ga. Selgub, et väga suurte valimite korral (kus s ja σ nagunii üsna sarnased tulevad) võib seda tõepoolest teha. Väiksemate valimite korral ($n < 60$) paraku nii toimida ei tohi. Lahendus on siiski olemas.

Kui juhuslik suurus $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ oli standardse normaaljaotusega juhuslik suurus, siis asendades populatsiooni standardhälbe σ valimi standardhälbega S saame (Studenti) t-jaotusega juhusliku suuruse:

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Hea sõnum seisneb selles, et t-jaotus on hästi teada, t-jaotuse kvantiilid on ära toodud igas endast lugupidavas statistikaalases raamatus ja seega kõik mis me ülaltoodud arutelus muutma peame, on see, et peame standardse normaaljaotuse kvantiilid asendama t-jaotuse vastavate kantiilidega. Halvem on see, et t-jaotuseid on palju. Iga valimi suuruse korral on meil tegemist teistest veidi erineva t-jaotusega. T-jaotuse määrab üheselt ära t-jaotuse parameeter - vabadusastmete arv (degrees of freedom — d.f.). Kvantiilid erinevate t-jaotuste tarvis on ära toodud tabelis 6.2.

Peale standardse normaaljaotuse kvantiilide asendamist t-jaotuse kvantiilidega jõuame järgmise usaldusintervalli arvutamise valemini. $(1-\alpha)$-usaldusintervall keskväärtusele on leitav järgmise valemiga:

$$\left(\bar{X} + t_{n-1;\alpha/2}\frac{S}{\sqrt{n}} \ldots \bar{X} + t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}}\right).$$

Näide 6.1 *Tuntud kanauurijat Hans Hane huvitab, mitu muna munevad keskmiselt eesti kanad ühe nädala jooksul. Kogumikus kanade kohta kavatseb ta esitada 95%-lise usaldusintervalli munade arvu keskväärtusele. Usaldusintervalli arvutamiseks luges härra Hani nädala jooksul kokku 10 kana munad: 5 5 4 6 1 7 5 6 3 3.*

Lahendus.

Kõigepealt leidis Hans valimi keskmise ja valimi standardhälbe: $\bar{x} = 4,5$*;* $s = 1,78$*. Seejärel leidis ta tabelist 6.2 arvutustes vajaminevad t-jaotuse kvantiilid:* $t_{0,025;9} = -2,26$ *ja* $t_{0,975;9} = 2,26$*. Asetades leitud arvud valemisse sai ta tulemuseks:*

$$\begin{aligned}
\left(\bar{x} + t_{n-1;\alpha/2}\frac{s}{\sqrt{n}} \quad \ldots \quad \bar{x} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\right) \\
\left(4,5 + (-2,26)\frac{1,78}{\sqrt{10}} \quad \ldots \quad 4,5 + 2,26\frac{1,78}{\sqrt{10}}\right) \\
(4,5 + (-2,26)0,56 \quad \ldots \quad 4,5 + 2,26 \times 0,56) \\
(3,23 \quad \ldots \quad 5,77)
\end{aligned}$$

Vastus: 95%-usaldusintervall eesti kanade nädala jooksul munetud munade arvu keskväärtusele on (3,23...5,77).

Kui soovime täpsemalt teada, milline tegelik keskväärtus olla võiks, tuleb suurendada valimi suurust. Mida suurem valim, seda kitsam tuleb ka usaldusintervall keskväärtusele.

6.3 Ülesanded

1. Uuriti sordi A saagikust. Keskmiseks saagikuseks saadi 1,2 tonni/ha; standardhälve oli 0,5. Milline on 95%-usaldusintervall sordi A saagikuse keskväärtusele, kui a) uuringu tulemused on saadud 8 põllu tulemuste põhjal (n=8) või b) uuringu tulemused on saadud 100 põllu tulemuste põhjal (n=100).

2. Kumb on laiem, kas 90% või 95%-usaldusintervall? Miks?

3. Kaks teadlast uurisid sordi B saagikust Eesti oludes. Nende kahe teadlase poolt leitud 95%-sed usaldusintervallid sordi B saagikuse keskväärtusele ei kattunud. Loetle põhjuseid, miks võidi samale küsimusele vastates saada mittekattuvad usaldusintervallid?

4. Üks teadlane mõõtis sordi C saagikust 8-l juhuslikult valitud põllul eestis. Teine teadlane mõõtis sordi C saagikust 8-l juhuslikult valitud katselapil ühel katsepõllul. Kumb teadlastest sai sordi C-saagikuse keskväärtusele laiema 95%-usaldusintervalli, miks? Kuidas tuleks kummagi teadlase poolt leitud usaldusintervalli interpreteerida?

Tabel 6.2: Studenti t-jaotuse kvantiilide t_α väärtused (nn kriitilised väärtused)

d.f. = n-1	$\alpha = 0,01$	$\alpha = 0,025$	$\alpha = 0,05$	$\alpha = 0,95$	$\alpha = 0,975$	$\alpha = 0,99$
1	-31,82	-12,71	-6,31	6,31	12,71	31,82
2	-6,97	-4,30	-2,92	2,92	4,30	6,97
3	-4,54	-3,18	-2,35	2,35	3,18	4,54
4	-3,75	-2,78	-2,13	2,13	2,78	3,75
5	-3,37	-2,57	-2,01	2,01	2,57	3,37
6	-3,14	-2,45	-1,94	1,94	2,45	3,14
7	-3,00	-2,36	-1,89	1,89	2,36	3,00
8	-2,90	-2,31	-1,86	1,86	2,31	2,90
9	-2,82	-2,26	-1,83	1,83	2,26	2,82
10	-2,76	-2,23	-1,81	1,81	2,23	2,76
12	-2,68	-2,18	-1,78	1,78	2,18	2,68
14	-2,62	-2,14	-1,76	1,76	2,14	2,62
16	-2,58	-2,12	-1,75	1,75	2,12	2,58
18	-2,55	-2,10	-1,73	1,73	2,10	2,55
20	-2,53	-2,09	-1,73	1,73	2,09	2,53
25	-2,49	-2,06	-1,71	1,71	2,06	2,49
30	-2,46	-2,04	-1,70	1,70	2,04	2,46
40	-2,42	-2,02	-1,68	1,68	2,02	2,42
60	-2,39	-2,00	-1,67	1,67	2,00	2,39
120	-2,36	-1,98	-1,66	1,66	1,98	2,36
∞	-2,33	-1,96	-1,64	1,64	1,96	2,33

5. 4-aastase kitse kaalu keskväärtus on 70kg, kaalu standardhälve on 3. Eeldades, et kitse kaal on normaaljaotusega juhuslik suurus (miks on see mõeldav eeldus?), leia 95%-prognoosintervall 4-aastaste kitsede kaalule.

6. uuritava tunnuse keskväärtus on 4, standardhälve on 1. Milline on 95%-usaldusintervall keskväärtusele?