Biomeetria bioloogidele

Statistiliste ja matemaatiliste meetodite
rakendamisest eluteadustes

Märt Möls
Matemaatilise Statistika Instituut
Tartu Ülikool

Euroopa Liit
Euroopa Sotsiaalfond

Eesti tuleviku heaks

2014

Sisukord

Sissejuhatus

Mis on biomeetria?

Kreeka keeles tähendab *bios* elu ja *metron* mõõtmist, seega on biomeetria elu mõõtmine. Biomeetria käsitleb matemaatika, statistika ja arvutustehnika rakendamist bioloogiliste probleemide lahendamiseks. Vahel kasutatakse ka termineid biostatistika (kui soovitakse esile tõsta statistikameetodite kasutamist) või bioinformaatika (kui tahetakse rõhutada arvutite ja arvutiteaduste osa). Käesolev kursus, olles kokku pandud statistiku poolt, kannab paratamatult oma looja jälge. Sestap on ta suuresti vaadeldav kui sissejuhatus biostatistikasse. Kitsamas tähenduses ongi biomeetria eelkõige statistiliste meetodite rakendamine bioloogiliste süsteemide uurimiseks.

Katkend Rahvusvahelise Biomeetriaseltsi koduleheküljelt (www.tibs.org):

The terms "Biometrics" and "Biometry" have been used since early in the 20th century to refer to the field of development of statistical and mathematical methods applicable to data analysis problems in the biological sciences. Statistical methods for the analysis of data from agricultural field experiments to compare the yields of different varieties of wheat, for the analysis of data from human clinical trials evaluating the relative effectiveness of competing therapies for disease, or for the analysis of data from environmental studies on the effects of air or water pollution on the appearance of human disease in a region or country are all examples of problems that would fall under the umbrella of "Biometrics".

Peatükk 1

Andmetest

Siin peatükis näeme, kui keerulise kõlaga sõnu saab kasutada andmetest rääkimisel; vaatame, kuidas algandmeid üles kirjutada ja kuuleme, milliseid silte võib mõõtmistulemustele külge kleepida

ehk

Andmemaatriksist, tunnuste tüüpidest ja kasutatavast tähistusest, väärtuste kodeerimisest ja puuduvatest väärtustest.

Selleks, et saaksime midagi objektiivset väita meid huvitava nähtuse, protsessi või objekti kohta on vaja vaatlus- või katsetulemusi. Enamasti pakub meile rohkem huvi see, milliseid järeldusi ja üldistusi me nende mõõtmistulemuste põhjal teha saame, kui üks või teine konkreetne mõõtmistulemus ise. Sestap on kerge unustada algmaterjal ja pürgida kohe kõrgustesse ja keerukate analüüside risträgastikku. Paraku ei seisa ükski maja kindlalt, kui vundament on hooletult ehitatud. Käesolevas peatükis vaatleme, millist terminoloogiat saab kasutada algandmetest rääkides ja sedagi, kuidas algandmeid nii kirja panna, et algandmetel tuginevad analüüsid hiljem ka kriitikatuultes püsima jääksid.

1.1 Objekt ja tunnus

Paljud statistikaga seotud arusaamatused on välditavad, kui inimesed saaksid täpselt aru, mida või keda nad uurivad (või keda on uuritud artiklis, mida parasjagu loetakse).

Definitsioon 1.1 ***Objekt*** *on uurimisalune ühik, üksikindiviid.*

Objektideks võivad näiteks olla linnupojad või pesakonnad; puud, metsatukad või punktid metsas; põdrad või ristamiskatse tagajärjel sündivad olevused.

Vahel on ka samade andmete puhul olemas mitu erinevat võimalust valida uurimisobjekti. Näiteks vaatame situatsiooni, kus on vaadeldud kahes pesas koorunud linnulapsi — ühes pesas koorus 9, teises 1 linnulast. Valides uurimisobjektiks pesakonna (kurna), saame tunnuse "*pesakonna suurus*" keskmiseks 5 (keskmine pesakonna suurus on 5); valides uurimisobjektiks aga linnulapse, saame tunnuse "*pesakonna suurus*" keskmiseks 8,2 (linnulaste keskmine päritolupesakonna suurus on 8,2). Vaata ka tabelit 1.1.

Tabel 1.1: Samad andmed - aga uurimisobjekt on erinev

Objekt - linnulaps

Pesakonna nr	Pesakonna suurus
1	9
1	9
1	9
1	9
1	9
1	9
1	9
1	9
1	9
2	1

Keskmine pesakonna suurus: 8,2

Objekt - pesakond

Pesakonna nr	Pesakonna suurus
1	9
2	1

Keskmine pesakonna suurus: 5

Näiteks metsa uurides võivad uurimisobjektideks olla puud või punktid metsas (võrdle: "80% vaadeldud metsapuudest olid pajud" vs "35% vaadeldud metsa-aladest olid kaetud pajudega"); taimede produktiivsuse uurimisel võib objektideks valida näiteks kas taime päevased või nädalased juurdekasvud (päeva või nädala jooksul lisanduv biomass), kusjuures uuritava tunnuse stabiilsus võib märgatavalt sõltuda meie valikust (nädala jooksul lisanduvad biomassid on sarnased; ühe päeva jooksul lisanduv biomass aga on üsna varieeruv suurus); jne.

Definitsioon 1.2 ***Tunnus*** *on objekti iseloomustav näitaja, mida põhimõtteliselt on võimalik mõõta või vaadelda.*

Hiiri uurides võivad tunnusteks olla karvavärv, kaal, sabapikkus, liik ja vanus; taimi uurides võivad tunnusteks osutuda kasvukoht, pikkus, lehtede arv, biomass, liik jne.

1.1.1 Tähistustest

Tunnuste nimede kirjutamisel kasutame edaspidi suuri tähti, näiteks $VANUS$, $SABAPIKKUS$ või $LIIK$. Vahel võime kasutada ka lühendeid, näiteks tunnuse "hiire poolt aasta jooksul läbinäritud raamatulehekülgede arv" võime tähistada sümboliga X (pane tähele - kasutame ikkagi suurt tähte X). Konkreetsete mõõdetud väärtuste tähistamiseks kasutame aga väikeseid tähti. Näiteks tunnuse X väärtus ühe konkreetse hiire puhul on tähistatud sümboliga x. Kui soovime täpsustada, millisel konkreetsel objektil vastav mõõtmine on aset leidnud, kasutame objekti numbrit alaindeksis: x_3 on tunnuse X väärtus 3. objektil (näiteks 3. hiire poolt rikutud lehekülgede arv).

Suurte tähtedega võime tähistada ka tulevasi mõõtmistulemusi, mille väärtus arutelu hetkeks pole selgunud. Näiteks planeerides järgmisel aastal aset leidvat uuringut saab rääkida 3. hiire mõõtmistulemusest kui suurusest X_3 (mis võib osutuda milleks iganes), peale uuringu toimumist ja andmete kogumist aga juba kui mõõtmistulemusest x_3 (mis on üks konkreetne ja teadaolev number).

1.2 Andmemaatriks

Arvuti jaoks andmete mõistetavaks tegemisel tuleb algandmed sisestada arvutisse kindlal kujul. Enamik statistikaprogramme soovib, et andmed oleksid sisestatud nn. objekt-tunnus maatriksina, st. sellise tabelina, kus iga veerg kujutab ühte tunnust ja iga rida ühte objekti.

Näide 1.1 *Käidi kümnel põllul ja koguti andmeid mullatüübi, mulla niiskuse ja viljakuse kohta. Saadud andmemaatriks on esitatud tabelis 1.2.*

Kaks võimalikku objekt-tunnus maatriksit on esinenud juba ka tabelis 1.1.

Tähtis on meeles pidada, et ühe (uurimis)objekti kohta tohib objekt-tunnus maatriksis olla vaid üksainus rida.

Kui andmed pole statistikaprogrammi sisestatud objekt-tunnus maatriksina, siis võib karta, et varem või hiljem leiab aset inimlik eksimus ning

Tabel 1.2: Objekt-tunnus maatriks

Põld	Mullatüüp	niiskus	suvinisu viljakus (kg/ha)
1	savimuld	niiske	3624
2	liivsavimuld	paras	4782
3	savimuld	niiske	4274
4	liivmuld	kuiv	3927
5	savimuld	paras	4630
6	liivmuld	paras	4920
7	savimuld	niiske	4260
8	savimuld	paras	4935
9	liivsavimuld	paras	5035
10	liivmuld	kuiv	4500

analüüsi tegev inimene interpreteerib arvuti poolt väljastatavaid tulemusi valesti.

Märkus: vahel kasutatakse ka andmemaatrikseid, kus ühe objekti kohta on kirjas mitu rida (nn kordusmõõtmiseid sisaldavad andmestikud). Selliste andmemaatriksite analüüs nõuab eriliste statistiliste meetodite kasutamist (kordusmõõtmiste analüüs, *repeated measures analysis*, ...) ja isegi pealtnäha lihtsatele küsimustele vastamine (milline on keskmine?) võib õige vastuse leidmine osutuda vägagi keeruliseks ülesandeks. Sellisel kujul esitatud andmete analüüs nõuab suuri statistika-alaseid teadmiseid ja pole enamasti algajale jõukohane.

1.3 Tunnuste tüübid

Võimalikke tunnuseid, mille vastu uurijad võivad huvi tunda, on sadu ja tuhandeid. Tunnuse uurimiseks sobivat metoodikat pole tarvis iga tunnuse jaoks uuesti leiutada – on ju võimalik leida keskmist nii hinnetele, saagikusele kui pesakonna suurusele. Kõigi mainitud tunnuste korral sobib keskmise arvutamiseks sama arvutuseeskiri.

Samas pole võimalik leida mullatüüpide keskmist – sellisel näitajal lihtsalt puuduks tähendus.

Kas oleks võimalik jagada tunnuseid selliselt, et ühte gruppi sattunud tunnused (sama tüüpi tunnused) on analüüsitavad kasutades sarnaseid statistikameetodeid? Selgub, et tunnuste taoline jagamine on täiesti võimalik.

Definitsioon 1.3 ***Pideva tunnuse*** *võimalike väärtuste arv on lõpmatu ja iga kahe võimaliku pideva tunnuse väärtuse vahele mahub alati veel üks võimalik pideva tunnuse väärtus.*

Pidevad tunnused on näiteks taime pikkus, looma kaal, temperatuur, fosfaatide kontsentratsioon vees, saagikus, ... Oleks soovitav, et kõik pideva tunnuse väärtused oleksid mõõdetud sama täpsusega (kõik pikkused mõõdetud millimeetri täpsuseni, kõik kaalumistulemused kirja pandud kg täpsusega jne). Igal juhul tuleb aga jälgida, et sama tunnuse kõik väärtused oleksid kirja pandud samades ühikutes (näiteks kilogrammides). Kui ühe elevandi kaaluna läheb kirja number 5300 (kg) ja teise elevandi kaaluks tuleb 4,9 (tonni), siis vaadeldud elevantide keskmiseks kaaluks annaks arvuti 2602,45, millisel numbril muidugi puudub igasugune sisu.

Definitsioon 1.4 ***Diskreetse tunnuse*** *väärtused saavad olla vaid täisarvulised. Peaaegu alati on diskreetse tunnuse väärtused tekkinud millegi loendamisel.*

Diskreetsed tunnused on näiteks pesakonna suurus, terade arv viljapeas, liikide arv ruutmeetril, looma poolt elu jooksul sünnitatud laste arv, ...

Definitsioon 1.5 ***Järjestustunnus*** *on tunnus, mille kõik võimalikud väärtused on järjestatavad.*

Järjestustunnused on näiteks eksperdi hinnang mullaniiskusele (väga kuiv - kuiv - paras - niiske - liigniiske); eksperdi hinnang looduskooslusele (riikliku kaitse alla võtta/ kohaliku kaitse alla võtta/ jätta juhuse hooleks/ buldooseritega hävitada); jälgija hinnang looma agressiivsusele (õel/ kurjavõitu/ normaalne/ rahumeelne/ tuim); aga samuti näiteks haridus, mõõdetuna skaalal algharidus - keskharidus - kõrgharidus - doktorikraad; jne.

Järjestustunnused tekivad sageli subjektiivsete hinnangute andmisel. Hinnangu andmise kriteeriumid võivad hindajati tugevalt erineda ning see võib paratamatult raskendada ka tulemuste hilisemat interpreteerimist. Seetõttu oleks tungivalt soovitav, et kõik hindajad ja ka analüüsi tulemuste hilisemad kasutajad mõistaksid võimalikult ühtemoodi seda, millal mulda on näiteks peetud "väga kuivaks" või kuna peeti kutsut õelaks.

Definitsioon 1.6 ***Nominaalne tunnus*** *on tunnus, mille väärtused pole sisuliselt järjestatavad.*

Nominaalsed tunnused on näiteks sugu, kasvukoht, liik, karvavärvus, lemmikroog, ...

Juhul, kui (nominaalsel) tunnusel on vaid kaks võimalikku väärtust, näiteks nagu tunnusel *SUGU*, siis kutsutakse vastavat tunnust ka **binaarseks** või **dihhotoomseks** tunnuseks.

Pidevaid ja diskreetseid tunnuseid kutsutakse vahel ka arvulisteks (kvantitatiivseteks) tunnusteks ja järjestus- ning nominaalseid tunnuseid kutsutakse mittearvulisteks ehk kvalitatiivseteks tunnusteks.

Näide 1.2 *Igal uuringusse kaasatud liblikal paluti ühe päeva jooksul muneda nii palju mune kui ta jaksab. Liblikat ja tema poolt munetud munasid uuriti põhjalikult. Kogutud andmed on esitatud tabelis 1.3. Märkus: toodud andmed on illustratiivsed ja ei baseeru tegelikel mõõtmistulemustel.*

Tunnused A ja C on pidevad, B on diskreetne, D on järjestustunnus (kasutatud kodeering: 1-väga räbaldunud; 2-kulunud; 3-peaaegu uus; 4-veatu), E on nominaalne (kasutatud kodeering: 1-rohetäpik; 2-suur-pärlmuttertäpik; 3-väike-pärlmuttertäpik).

Tabel 1.3: Liblikad

A liblika suurus	B munade arv	C munetud munade keskmine kaal	D liblika ilu	E liik
11	20	1,3	2	1
10	34	1,8	3	1
12	67	0,9	4	2
7	10	0,7	2	3
12	0	1,0	1	2

1.4 Tunnuste kodeerimine, puuduvad väärtused

Tunnuse väärtuste sisestamisel arvutisse on sageli mõistlik üks või teine (enamasti pikk) väärtus asendada lühendi ehk koodiga. Näiteks tunnuse *PÜÜGIKOHT* väärtuste sisestamisel võime väärtuse "Elva-Vitipalu maastikukaitseala" asemel sisestada numbri "1" jne. Järjestustunnuse puhul on tunnuse väärtuste kodeerimine numbrite abil enamasti tungivalt soovitav. Sealjuures tuleks jälgida, et koodid säilitaksid väärtuste sisulise järjestuse. Seega ei tohi kasutada kodeeringut:

```
1 - hea
2 - paha
3 - ei oska öelda
```

küll aga sobivad kodeeringud

```
1 - hea             1 - paha             1 - hea
2 - ei oska öelda   2 - ei oska öelda    0 - ei oska öelda
3 - paha            3 - hea             -1 - paha
```

Iga veidigi suurema uuringu paratamatuks kaaslaseks on puuduvad andmed. Kas keeldub mõni talunik vastamast mõnele küsimusele, unustas doktorant katselapil ettenähtud ajal mõõtmisi tegemas käia või lasi katses kasutatud valge hiir lihtsalt jalga — tegijal juhtub nii mõndagi. Puuduvad andmed võivad analüüsi käigus palju peavalu põhjustada. Sellegi poolest tasub meeles pidada, et puuduvate andmete lihtsalt "äraunustamine" pole enamasti lahendus ja tekitab tavaliselt rohkem probleeme kui lahendab. Sestap on tungivalt soovitav algandmete sisestamisel sisestada ka need kirjed/objektid, kelle kohta andmed (osaliselt) puuduvad. **Puuduvad väärtused peavad andmestikus olema tähistatud nii, nagu ei tähistata andmestikus midagi muud.** Eriti kergesti võivad puuduva väärtusega segi minna näiteks tegelikud mõõdetud väärtused "0" või "vaadeldud omadust ei esinenud".

Näiteks parasiit A olemasolu mõõtev tunnus võiks olla kodeeritud järgmiselt: "1" – parasiit esines; "0" – parasiiti polnud; "." – informatsioon puudub (puuduv väärtus).

1.5 Ülesanded

1. Uuringu käigus koguti andmeid 15 jahimeeste poolt lastud põdra kohta. Kogutud andmed on esitatud tabelis 1.4. Milliseid tunnuseid mõõdeti? Mis tüüpi tunnustega on tegemist? Milline näeks välja objekt-tunnus maatriks antud andmete korral?

2. Loomaembrüodel mõõdeti järgmiste tunnuste väärtused:

   - VANUS1 (vanus päevades)
   - VANUS2 (rakkude pooldumiskordade arv)
   - EMASLOOMA EKSPOSITSIOON ALKOHOLILE (ei/ natuke/ ohtralt)
   - GEENI X MUTATSIOON (esines/ ei esinenud)

Tabel 1.4: Ülesanne 1 - Kolmes metsas lastud põtrade vanused

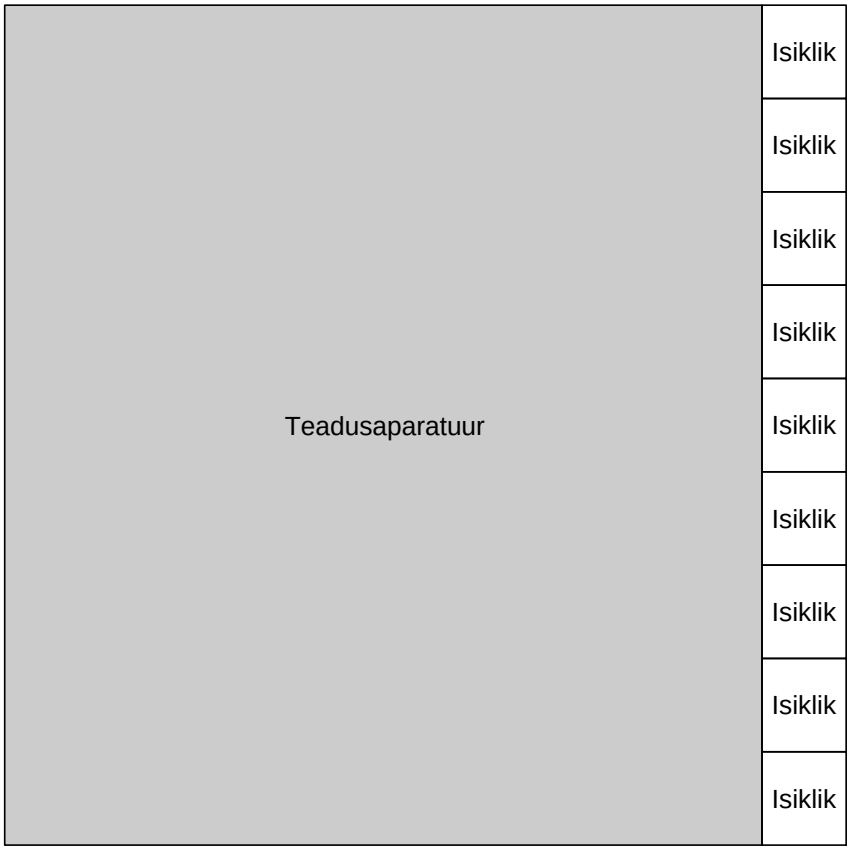
	laskmiskoht		
aasta	Alutaguse	Vändra kant	Kapa-Kohila
2004	10a, 12a	3a	
2005	10a	3a, 5a	7a, 15a, 10a
2006	10a, 11a	4a, 15a	4a, 8a

- KASVUKESKONNA Ph

Mis tüüpi tunnustega on tegemist?

3. Ajakirjanduses ilmusid väited, et Antarktikasse rajatud Eesti uurimisjaama maksumaksja kulul saadetud asjadest on 90% loodusuurijate isiklikud asjad. Loodusuurijad vaidlesid vastu, et isiklikud asjad moodustasid kõnealusest saadetisest vaid 10%. Milles on asi? Kellel on õigus? Vaata ka joonist 1.1!

Joonis 1.1: Saadetis Antarktikasse



Peatükk 2

Kirjeldav statistika

Siin peatükis kuuleme, kuidas saaks lühidalt ja kokkuvõtlikult kirjeldada äraütlemata suurt lasu kokkukogutud andmeid
ehk
ühe tunnuse empiirilise jaotuse kirjeldamine.

2.1 Sagedused ja osakaalud

Kõige lihtsam ja ülevaatlikum viis andmeid kirjeldada, eriti kui tegemist on nominaalse, järjestus- või väheste võimalike väärtustega diskreetse tunnusega on iga võimaliku väärtuse kohta öelda, mitu korda me sellist väärtust oleme näinud (raporteerida iga väärtuse esinemissagedust). Enamasti esitatakse selline kokkuvõtlik informatsioon kas sagedustabelina, tulp- või ringdiagrammina. Vahel on otstarbekam välja tuua erinevate väärtuste osakaalud — näiteks kui suur osa kõigist põldudest asuvad liivastel muldadel, kui suur osa savimuldadel jne. Osakaalusid võib vajaduse korral esitada ka protsentides.

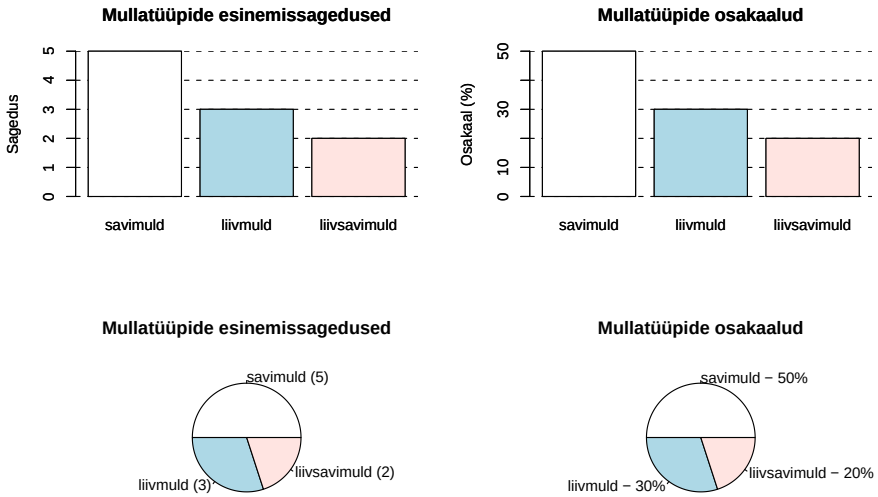
Alljärgnevalt vaatleme näites 1.1 toodud tunnuse *MULLATÜÜP* sagedus- ja jaotustabelit ning nende tabelite põhjal joonistatud tulp- ja ringdiagramme.

Nominaalsete tunnuste väärtuste kirjeldamisel polegi suurt midagi muud võimalik teha, kui esitada tunnuse sagedustabel (või esitada erinevate väärtuste osakaalud). Vahel tuuakse eraldi välja ka tunnuse mood — kõige sagedamini esinenud tunnuse väärtus. Mullatüüpe iseloomustavas näites oleks tunnuse *MULLATÜÜP* moodiks savimuld - savimuldasid esines vaadeldud põldude seas kõige rohkem.

Pideva tunnuse väärtuste iseloomustamisel pole ülaltoodud viisil koos-

Tabel 2.1: Tunnuse *MULLATÜÜP* võimalike väärtuste sagedused ja osakaalud

Mullatüüp	Sagedus	Osakaal	Osakaal (%)
savimuld	5	0,5	50%
liivmuld	3	0,3	30%
liivsavimuld	2	0,2	20%



tatud sagedustabelist aga eriti abi — tabel tuleks liiga pikk ning poleks märkimisväärselt targem kui lihtsalt kõigi tunnuse vaadeldud väärtuste esitamine. Koostamaks mõistlikku sagedustabelit pideva tunnuse tarvis, jagatakse pideva tunnuse väärtused eelnevalt samapikkadeks vahemikeks (Näiteks [0..10),[10..20), jne). Seejärel vaadatakse, kui sageli pideva tunnuse väärtus sattub ühte või teise vaatlusalusesse vahemikku. Mitut vahemikku kasutada? Kindlat reeglit siin pole. Üks soovitus võiks olla järgmine: leia ruutjuur vaatluste arvust. Vali vahemike arvuks mõni täisarv, mis oleks ligikaudu samasuur kui leitud ruutjuur. Näiteks, kui tehtud on 10 vaatlust, siis võiks pideva tunnuse väärtused jagada 3 või 4 vahemikku. Antud reegel on vaid soovitusliku väärtusega, vajadusel võib kasutada ka rohkemaid või vähemaid

vahemikke.

Pideva tunnuse sagedustabeli illustreerimisel kasutatavat joonist kutsutakse histogrammiks. Kui tulpdiagramm oli antud andmete (vaatlustulemuste) korral üheselt määratud, siis samade andmete põhjal võime saada üsnagi erineva kujuga histogramme. Muutes sagedustabeli koostamisel kasutatud vahemikke muutub enamasti ka sagedustabel ja tema põhjal joonistatud histogrammi kuju.

Näide 2.1 *Näites 1.1 on antud suvenisu viljakused erinevatel põldudel. Suvenisu viljakus on pidev tunnus, tema väärtuste jaoks sagedustabeli koostamisel võiksime jagada viljakusandmed näiteks nelja vahemikku. Saadud sagedustabel on antud tabelis 2.2. Antud andmete illustreeriva histogrammi allosas on ära märgitud väikeste joonekestega ka tegelikud vaatlusandmed. Kasutatud vahemike korrektsel kirjeldamisel võib kasutada ka nurk- ja ümarsulge – piiripeale jääv vaatlus pannakse siis kirja sinna vahemikku, kus vastav väärtus piirneb nurksuluga. Seega mõõtmistulemus* 4500 *läheb kirja vahemikku* $[4500 \ldots 5000)$ *ja mitte vahemikku* $[4000 \ldots 4500)$.

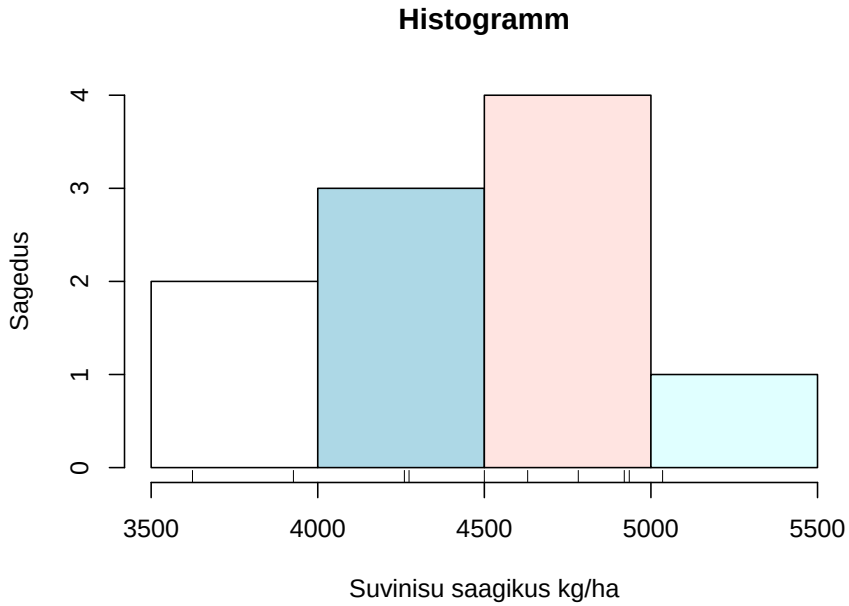
Tabel 2.2: Tunnuse *VILJAKUS* sagedustabel ja osakaalud

Viljakus	Sagedus	Osakaal	Osakaal (%)
$[3500 \ldots 4000)$	2	0,2	20%
$[4000 \ldots 4500)$	3	0,3	30%
$[4500 \ldots 5000)$	4	0,4	40%
$[5000 \ldots 5500)$	1	0,1	10%

NB! On tungivalt soovitav, et kõik kasutatud vahemikud oleksid võrdse pikkusega! Sestap tuleks võimaluse korral vältida ka "avatud" vahemikke, nagu näiteks "suurem kui 50 ha". Kasutades muutuva pikkusega vahemikke võib statistika tarbijas tekitada just sellise pettekujutelma nagu keegi parasjagu soovib.

Näide 2.2 *Kasutades muutuva pikkusega vahemikke sagedustabeli koostamisel võib hooletut või kehva ettevalmistusega statistika tarbijat petta andmeid võltsimatta just nii, nagu parasjagu tarvis. Graafikul 2.2 on samad pideva tunnuse väärtused esitatud kahel erineval moel - kasutades võrdse pikkusega vahemikke sagedustabeli koostamisel (soovitatav tegutsemisviis) ja kasutades muutuva pikkusega vahemikke (petturlus pole haritud inimesele sobiv tegutsemisviis).*

Joonis 2.1: Näites 2.1 toodud sagedustabeli põhjal joonistatud histogramm

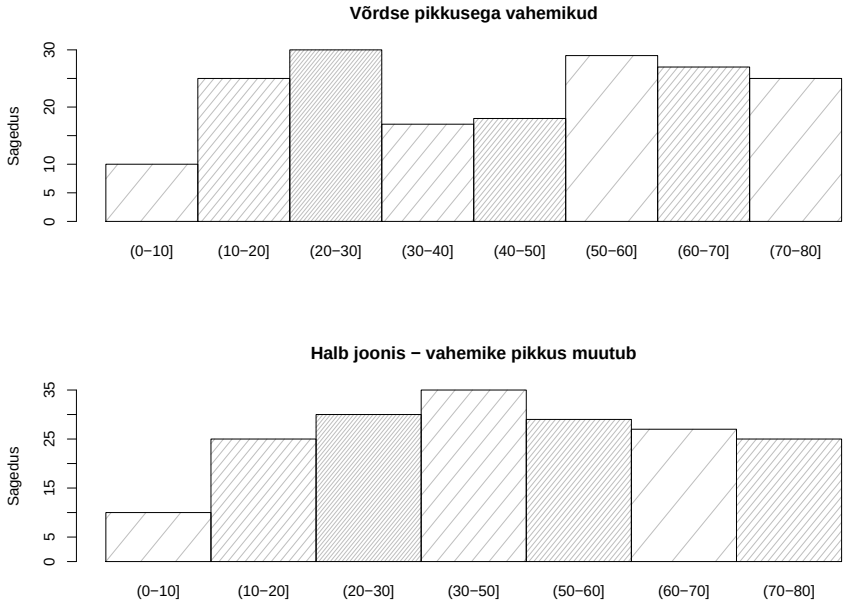


NB! Joonisele tuleb kanda ka vahemikud, kuhu ühtki objekti ei sattunud (kui mingisse vahemikku ei sattunud ühtegi objekti, jätavad paljud programmid sagedustabeli koostamisel vastava rea tabelist välja ja eksitus graafiku joonistamisel on siis juba kerge tulema)!

2.2 Statistikud

Sagedustabel on kasulik viis uuritud tunnuse väärtuste iseloomustamiseks, aga vahel soovime tähelepanu juhtida mõnele meid kõige enam huvitavale andmetega seotud küljele või rõhutatult välja tuua meie andmete omapära. Sealjuures on meile tihti abiks statistikud. *Statistik* on andmete põhjal üheselt arvutatav (enamasti arvuline) näitaja. Arvatavasti tuntuim statistik on keskmine (kõik me oleme muretsenud oma keskmise hinde pärast või lugenud ajalehtedest keskmise palga muutumisest).

Joonis 2.2: Samad andmed - võrdse pikkusega vahemikud ja muutuva pikkusega vahemikud



2.3 Vaatlustulemuste suurust iseloomustavad statistikud

2.3.1 Keskmine

Inglise keeles *mean* või *average*, $\bar{x}$ - uuritava tunnuse väärtuste aritmeetiline keskmine:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n\right).$$

Näide 2.3 *Eesrindlik talunik Sauna Mats hakkas oma tiigis krokodille kasvatama. Kasvatas neid veidi ja mõõtis siis kõigi oma kuue kasvandiku pikku-*

sed ära: 2,3m 1,7m 0,6m 0,8m 1,4m 2,2m. Nende keskmine pikkus on seega

$$\begin{aligned}\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}\left(x_1 + x_2 + \ldots + x_n\right) \\ &= \frac{1}{6}\left(2{,}3 + 1{,}7 + 0{,}6 + 0{,}8 + 1{,}4 + 2{,}2\right) = \frac{1}{6}9 = 1{,}5\,(m).\end{aligned}$$

Keskmise omadusi:

1. $\overline{cx} = c\bar{x}$, kus c on konstant.

   Tõestus:

   $$\overline{cx} = \frac{1}{n}\sum_{i=1}^{n} cx_i = c \cdot \frac{1}{n}\sum_{i=1}^{n} x_i = c \cdot \overline{x}.$$

   Üks järeldus sellest omadusest: Kui oleksime näiteks samade objektide pikkuseid mõõtnud meetrites (x) ja sentimeetrites ($100x$), siis sentimeetrites tehtud mõõtmiste keskmine ($\overline{100x}$) tuleks sama kui meetrites tehtud mõõtmiste keskmine ($\bar{x}$) korda 100.

2. $\overline{x + c} = \bar{x} + c$, kus c on konstant.

   Tõestus:

   $$\overline{x + c} = \frac{1}{n}\sum_{i=1}^{n}(x_i + c) = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} c = \overline{x} + c.$$

   Kui näiteks peale püütud loomade kaalumist selgus, et kaal polnud täpselt tasakaalus - igal kaalumisel näitas kaal c kilogrammi rohkem, siis valede kaalumiste keskmist $\overline{x + c}$ teades saame arvutada korrektse kaaluga tehtud kaalumiste keskmise: $\bar{x} = \overline{x + c} - c$ ning seega pole keskmise arvutamiseks tarvis mõõtmiseid uuesti teha.

3. $\overline{x + y} = \bar{x} + \bar{y}$

   Tõestus:

$$\begin{aligned}\overline{x+y} &= \frac{1}{n}\sum_{i=1}^{n}(x_i+y_i) = \frac{1}{n}(x_1+x_2+\ldots+x_n+y_1+y_2+\ldots+y_n)\\ &= \frac{1}{n}\sum_{i=1}^{n}x_i + \frac{1}{n}\sum_{i=1}^{n}y_i = \overline{x}+\overline{y}.\end{aligned}$$

Üks tudeng arvutas, kui palju kasvavad keskmiselt taimekesed hommiku ja lõuna jooksul ($\bar{x}$). Teine tudeng leidis õhtu ja öö jooksul toimunud juurdekasvude keskmise ($\bar{y}$). Professor leidis aga taimede keskmise ööpäevase juurdekasvu ($\overline{x+y}$) liites oma kahe tudengi tulemused ning kirjutas selle kohta artikli ise ühtegi mõõtmist tegemata.

4. $\sum_{i=1}^{n} x_i = n\bar{x}$
   Triviaalne, kuid igal juhul meelespidamist vääriv. Näiteks karja summaarne väljalüps on leitav korrutades keskmise väljalüpsi karja suurusega.

Binaarse (kahe võimaliku väärtusega) tunnuse keskmine väärib eraldi ära märkimist. Kui meil on näiteks tegemist tunnusega võimalike väärtustega 0 (parasiite pole) ja 1 (parasiite leidub, n_1 vaatlust), siis taolise tunnuse keskmiseks tuleb 1-tede osakaal (või, korrutatult sajaga, protsent):

$$\overline{x} = \frac{1}{n}(1+0+0+1+0+\ldots) = \frac{n_1}{n}.$$

Seega on taolisel viisil kodeeritud andmete korral keskmise leidmine kerge viis parasiitidega nakatunud isendite protsendi arvutamiseks.

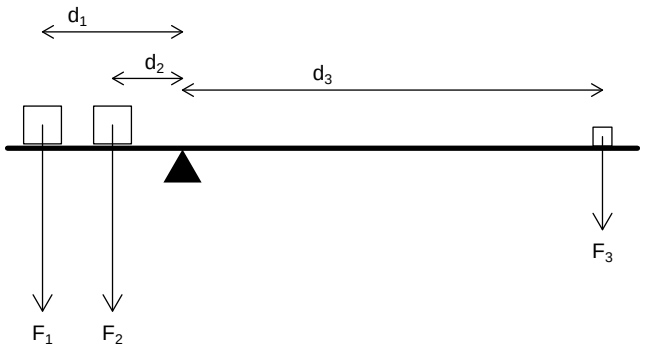
Üks võimalus valimi keskmist paremini tunnetuslikult tajuda on kasutades kangkaalu näidet. Kujutame nüüd endale ette arvtelge, kus tugi on pandud valimikeskmise $\overline{x}$ alla. Laome sellele arvteljele oma vaatlused, kõik kaalult võrdsed. Selgub, et selline kangkaal jääb tasakaalu, kui paigutame tema toetuspunkti valimikeskmise $\overline{x}$ alla.

Toodud väite põhjendus. Vaata joonisel 2.3.1 kujutatud kangkaalu.

Joonisel kujutatud kangkaal püsib tasakaalus kui $F_1 \cdot d_1 + F_2 \cdot d_2 = F_3 \cdot d_3$. Üldisemal kujul kirja pandult: kangkaal püsib tasakaalus, kui mõlemale kangi õlale mõjuvad samasuured jõumomendid: $\sum_{j\in\text{vasak pool}} F_j d_j = \sum_{i\in\text{parem pool}} F_i d_i$.

Valimikeskmisest suurema vaatluse ($x_i > \overline{x}$) kaugus tugipunktist ehk valimikeskmisest on $d_i := x_i - \overline{x}$ ja nende summaarne jõumoment (millega nad kangi paremalt poolt alla suruvad) on $\sum_{x_i>\overline{x}} 1 \cdot d_i = \sum_{x_i>\overline{x}}(x_i - \overline{x})$.

Joonis 2.3: Kangkaal.

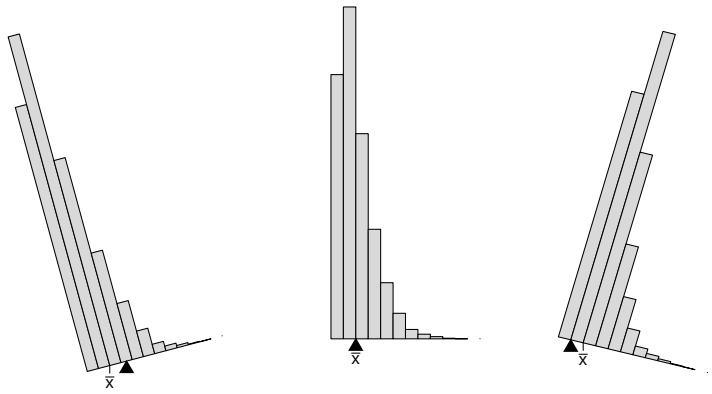


Valimikeskmisest väiksema vaatluse ($x_i < \overline{x}$) kaugus tugipunktist on aga $\overline{x} - x_i$ ja selliste vaatluste summarne jõumoment (millega nad kangi vasakult poolt alla suruvad) on $\sum_{x_i<\overline{x}} 1 \cdot d_i = -\sum_{x_i<\overline{x}}(x_i - \overline{x})$. Selgub aga, et antud juhul ongi mõlemale kaalu õlale mõjuvad jõumomendid võrdsed:

$$
\begin{aligned}
\overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
0 &= \frac{1}{n}\left(\left(\sum_{i=1}^{n} x_i\right) - n\overline{x}\right) \\
0 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) \\
0 &= \sum_{x_i<\overline{x}}(x_i - \overline{x}) + \sum_{x_i\geq\overline{x}}(x_i - \overline{x}) \\
-\sum_{x_i<\overline{x}}(x_i - \overline{x}) &= \sum_{x_i\geq\overline{x}}(x_i - \overline{x}) \\
-\sum_{x_i<\overline{x}}(x_i - \overline{x}) &= \sum_{x_j>\overline{x}}(x_i - \overline{x}) + \sum_{x_j=\overline{x}}(x_i - \overline{x}) \\
-\sum_{x_i<\overline{x}}(x_i - \overline{x}) &= \sum_{x_i>\overline{x}}(x_i - \overline{x}) + 0
\end{aligned}
$$

Teades sarnasust kangkaaluga võime ka kergesti lugeda histogrammi pealt välja valimikeskmise ligikaudse väärtuse. Hindame lihtsalt silma järgi koha, kuhu tuleks histogrammi alla paigutada toetuspunkt, nii et histogramm jääks tasakaalu ega vajuks ei paremale ega vasakule poole kaldu, vaata näiteks joonist 2.3.1.

Joonis 2.4: Histogramm ja valimi keskmine



Keskmine pole alati parim näitaja iseloomustamaks uuritava tunnuse väärtuste suurust. Nimelt on aritmeetiline keskmine tundlik üksikute suurte väärtuste suhtes - piisab ühestainsast teistest märgatavalt erinevast vaatlusest, et keskmist tugevalt muuta. Seda illustreerib ka järgmine näide.

Näide 2.4 *Uurides maduusside kaalu, saadi vaatlustulemusteks 2,2kg 2,6kg 2,8kg 2,4kg 10,0kg. Viimane madu on sedavõrd kaalukas, kuna on vahetult enne ülekaalumist nahka pistnud saaklooma.*

$$\bar{x} = \frac{1}{5}\left(2{,}2 + 2{,}6 + 2{,}8 + 2{,}4 + 10{,}0\right) = 20/5 = 4\,(kg).$$

Selgub, et keskmine kaal tuleb suurem kui enamike usside kaal ja peegeldab tugevalt ebatüüpilise, äsjatoitunud looma kaalu. Kuna keskmine võib olla suurem (või väiksem) kui enamik vaatlustulemusi, tekib lahknevus aritmeetilise keskmise kui näitaja ja keskmise intuitiivse tähenduse vahel (intuitiivselt on ju selge, et "keskmine" madu kaalub vähem kui 4 kg!!). Pakkumaks välja teist matemaatilist näitajat, mis iseloomustab andmete "keskmist" suurust sageli intuitiivselt täpsemalt, on kasutusele võetud mediaan.

2.3.2 Mediaan

inglise keeles *median* (või lühendina *med*) - vaatlustulemus, millest suuremaid ja väiksemaid väärtuseid on samapalju. Mediaani leidmiseks järjestatakse kõik vaatlustulemused, saades nn. variatsioonrea: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, kus $x_{(1)}$ on kõige väiksem vaatlustulemus, $x_{(2)}$ on suurem kui $x_{(1)}$ kuid väiksem kõigist ülejäänutest jne kuni vaatlustulemuseni $x_{(n)}$, mis on suurem kõigist teistest. Selle järjestatud vaatlustulemustest moodustatud rea keskel asuv vaatlustulemus ongi mediaan. Kui keskmist vaatlustulemust ei saa leida (kui vaatlustulemusi on paarisarv tükki) siis sobib mediaaniks mistahes arv kahe variatsioonrea keskmise elemendi vahel. Kokkuleppeliselt loetakse sellisel juhul mediaaniks kahe variatsioonrea keskel asuva vaatlustulemuse aritmeetilist keskmist. Matemaatiliselt korrektselt kirjapandult: Kui vaatlusi on tehtud paaritu arv kordi, $n = 2k+1$, siis on vaatlustulemuste mediaaniks variatsioonrea $(k+1)$. element $x_{(k+1)}$. Kui vaatlustulemusi oli aga paarisarv, $n = 2k$, siis loetakse vaatluste mediaaniks variatsioonrea k. ja $(k+1)$. elemendi aritmeetilist keskmist: $med(X) = (x_{(k)} + x_{(k+1)})/2$. Variatsioonrea j-ndat elementi, $x_{(j)}$, nimetatakse j.-ndaks järkstatistikuks. Vaatlustulemuse järjekorranumbrit variatsioonreas nimetatakse astakuks.

Näide 2.5 *Leiame eelmises näites toodud maduusside kaalu mediaani. Esmalt järjestame vaatlustulemused leidmaks variatsioonrida ja saame: 2,2 2,4 2,6 2,8 10,0. Kuna vaatlusi on paaritu arv,* $n = 5 = 2 \cdot 2 + 1, k = 2$, *siis saame mediaaniks variatsioonrea 3. elemendi* $med(X) = x_{(2+1)} = x_{(3)} = 2{,}6$. *Tulemus iseloomustab maduusside harilikku kaalu paremini kui keskmine kaal, sest on vähem mõjutatav eriliste üksikisendite (üksikvaatluste) poolt.*

Mediaani omadustest:

erinevatel põhjustel vaatlusandmeid vahel teisendatakse, näiteks logaritmitakse. Juhul, kui kasutatav teisendus ei muuda vaatluste järjekorda (suurim jääb suurimaks jne — või täpsemalt öeldes: kui kasutatav teisendus on monotoonne), siis võime teisendatud andmete mediaani leida tehes esialgsete andmete põhjal leitud mediaaniga sama teisenduse (näiteks logaritmime). Matemaatilisemas keeles öeldult: teostades mistahes andmete monotoonse teisenduse $f(x)$, st asendades vaatlusandmed $x_1, x_2, \ldots, x_n$ teisendatud vaatlusandmetega $x_1^{uus} = f(x_1), \ldots, x_n^{uus} = f(x_n)$ võime teisendatud andmete mediaani arvutada esialgsete andmete mediaani kasutades:

$$med(x^{uus}) = f(med(x)).$$

Samuti tuleks meeles pidada, et mediaani ja vaatluste arvu teades ei saa välja rehkendada vaatluste summat — mis on tuntav puudus. Riigi mediaan-

palka ja töötegijate arvu teades pole riigiametnikul või ärimehel võimalik leida summaarset palkadena ringlevat rahasummat; päevaste läbimüükide mediaani teades ei saa poodnik leida kuu summaarset läbimüüki jne.

Teatavatel juhtudel võib mediaan osutuda liiga "tuimaks" statistikuks: kuigi andmed muutuvad küllaltki palju, mediaan ei muutu.

Näide 2.6 *Soovime võrrelda linnamüras ja saastas elava linnupaari kurna suurust metsarahus pesitseva linnupaari omaga. Kogutud andmed on järgmised:*

Linnas pesitsevad linnud: 1, 1, 1, 2, 2, 2, 2

Metsas pesitsevad linnud: 2, 2, 2, 2, 3, 5, 7

Mõlemal juhul tuleb pesakonna suuruse mediaaniks 2 linnupoega. Seega neid kahte gruppi mediaani abil võrreldes me ei märkakski pesitsedukuse erinevust.

Kuna viimases näites ülestõstetud "liigse-tuimuse-probleem" esineb eelkõige järjestus- või diskreetsete tunnuste korral, tasub mediaani kasutada nimetatud tüüpi tunnuste korral üsna ettevaatlikult.

2.3.3 Mood

Inglise keeles *mode*, lühendina ka *mod* — vaatlustulemus, mida esineb kõige rohkem — väärtus, mis on parajasti moes.

Näide 2.7 *Uuriti metsatukas kasvavate seente liigilist koostist. Saadi tulemuseks: puravik, sitaseen, kukeseen, puravik, kukesen, kukeseen, sitaseen, kukeseen.*

mod(X)=kukeseen.

Arvuliste väärtustega tunnuste jaoks moodi leides võime sattuda olukorda, kus iga või enamik vaatlustest teineteisest erinevad (kui mõõta piisavalt täpselt, selgub, et iga maduussi kaal on teistest erinev). Enamasti kasutatakse siis sagedustabeli abi (vaata sagedustabeli koostamist pidevale tunnusele) ja vahemikku, mis osutus kõige populaarsemaks, loetakse moodiks. Sõltuvalt histogrammi kujust räägitakse vahel ka kahemodaalsest jaotusest (histogrammil on kaks eraldipaiknevat tippu), või multimodaalsest jaotusest (rohkem kui kaks eraldipaiknevat tippu).

Küsimus:
Eksperimendi keskmine kestvus on 4 päeva. Kas reserveerides endale aega eksperimendi läbiviimiseks 5 päeva, võime olla kindlad, et jõuame selle ajaga tulemusteni? Mida oleks vaja teada lisaks antud eksperimendi kestvuse kohta, et suudaksime sellele küsimusele vastata?

2.4 Vaatluste hajuvus

Mõnikord on kõik vaatlused igavalt üheülbalised, teinekord on aga iga uus mõõtmistulemus teistest sedavõrd erinev, nagu polekski mõõdetud ühe ja sama tunnuse väärtust. Kui erinevad teineteistest vaatlustulemused antud tunnuse korral võivad olla?

2.4.1 Miinimum ja maksimum

inglise keeles *maximum*, *minimum*, lühendina kui min, max — sageli kasutatavad ja intuitiivselt hästi mõistetavad tunnuse hajuvust (võimalikku varieeruvust) iseloomustavad statistikud. Teades näiteks eksperimendi minimaalset ja maksimaalset võimalikku kestvust, saaksime kindlalt väita, kas meie poolt eksperimendi jaoks planeeritud 5 päevast piisab.

Tunnuse hajuvuse iseloomustamiseks kasutatakse ka maksimumi ja miinimumi vahet ehk haaret (ka variatsiooniulatus, inglise keeles *range*) — maksimumi ja miinimumi vahe:

$$haare = maksimum - miinimum$$

Ehkki kergesti mõistetavad, esineb miinimumi ja maksimumi kasutamisel ka tõsiseid probleeme. Juhul, kui uuritavaks tunnuseks on jalgade arv küülikul, võime kergesti jõuda tulemuseni: jalgu on küülikul 0 (miinimum) kuni 8 (maksimum). Miks? Sest aeg-ajalt sünnib väärarengutega küülikuid, esineb vigastatud loomi jms. Korraliku uuringu ja ausa uurija puhul kirjeldavad miinimum ja maksimum sageli geneetiliselt muteerunud või muidu väga harukordseid ja erandlikke isendeid või juhtumeid. Tänu omadusele kirjeldada kõige veidramaid juhtumeid, võivad miinimum ja maksimum osutuda praktikas kehvasti kasutatavaks - suur osa vaatlustulemusi on enamasti märksa suuremad kui miinimum ja märksa väiksemad kui maksimum. Praktikas aeg-ajalt esinev lähenemisviis, kus uurija oma meele järgi suvaliselt "ebatüüpiliste" isendite mõõtmistulemused minema viskab enne miinimumi ja maksimumi leidmist, pole teaduskirjanduses lubatav - sest iga uurija jaoks

võib "ebatüüpiline" omada erinevat tähendust. See raskendab miinimumi ja maksimumi kasutamist uuritava tunnuse teaduslikul kirjeldamisel, teeb nad aga väärtuslikuks andmetest vigade või veidrike väljaotsimisel.

On ka teine probleem, mis on seotud miinimumi ja maksimumi kasutamisega. Uute mõõtmistulemuste selgumisel saab vaatlustulemuste maksimum ainult kasvada. Näiteks huvitagu meid küülikute kaal. Mõõtnud ära saja või tuhande küüliku kaalud, võib ta ikkagi olla üsna kindel, et kuskil lippab veelgi priskem isend. Sealjuures on üsna võimatu olemasolevate andmete põhjal oletada, kui kaalukas võib olla Eesti Kõige Kaalukam Küülik.

Miinimum ja maksimum on üks võimalus iseloomustada tunnuse hajuvust. Teine võimalus on iseloomustada hajuvust kirjeldades üksikvaatluste kaugust keskmisest. Seda ideed modifitseerides on saadud dispersiooni nime all tuntud statistik.

2.4.2 Dispersioon ja standardhälve

Mõiste dispersiooni vaste inglise keeles on *variance*, tähistus: s^2. Dispersiooni arvutamiseks kasutatakse järgmist valemit:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Üksikvaatluste erinevus keskmisest, $x - \overline{x}$, nimetatakse hälbeks. Dispersiooni saab vaadata kui hälvete ruutude keskmist. Miks dispersiooni arvutamisel keskmise leidmiseks kasutatakse jagajana suurust $(n-1)$ tavalise n-i asemel, sellel peatume lähemalt järgmises loengus.

Kui kõik vaatlused on samasuured (kõigil uuritavatel loomadel on neli jalga), siis on kõik hälbed keskmisest nullid ja uuritava suuruse dispersioon on null (tunnuse "jalgade arv" dispersioon on null). Mida erinevamad keskmisest on vaatlused, seda suurem on ka dispersioon.

valimi standardhälve (s) - inglise keeles *standard deviance*, lühendina ka *sd* või *std*. On samuti tunnuse hajuvust kirjeldav näitaja,

$$s = \sqrt{s^2}.$$

Sarnane dispersioonile, kuid on teisendatud viimaks mõõtühikuid võrreldavaks uuritava tunnuse algsete ühikutega. Tunnetuslikult tajutav kui vaatluste teatavat sorti keskmine kaugus keskmisest.

Näide 2.8 *Uuriti kahte gruppi hiiri: metsikuid ja geneetiliselt puhtatõulisi laborihiiri. Mõlemas grupis mõõdeti hiirte reaktsiooni ärritajale. Tulemuseks saadi:*

```
Metsikud hiired: 15, 45, 30, 10, 25
Laborihiired:    20, 25, 30, 25
```

Standardhälbe leidmiseks tuleb esmalt leida mõlema grupi jaoks keskmised:

$$\overline{x_{metsik}} = \frac{1}{5} \times (15 + 45 + 30 + 10 + 25) = 25$$

$$\overline{x_{labor}} = 25$$

Leiame valimi dispersioonid:

$$\begin{aligned} {s_{metsik}}^2 &= \frac{1}{4}((15-25)^2 + (45-25)^2 + (30-25)^2 + (10-25)^2 + (25-25)^2) \\ &= \frac{1}{4}(100 + 400 + 25 + 225 + 0) = 750/4 = 187{,}5 \\ {s_{labor}}^2 &= \frac{1}{3}((20-25)^2 + (25-25)^2 + (30-25)^2 + (25-25)^2) \\ &= \frac{1}{3}(25 + 0 + 25 + 0) = 50/3 = 16{,}66... \end{aligned}$$

Kust saame juba valimi standardhälbed mõlema grupi jaoks:

$$\begin{aligned} s_{metsik} &= \sqrt{{s_{metsik}}^2} = \sqrt{187{,}5} = 13{,}69\ldots \\ s_{labor} &= \sqrt{{s_{labor}}^2} = \sqrt{16{,}66\ldots} = 4{,}08\ldots. \end{aligned}$$

Märkame, et laborihiired reageerivad ärritusele märksa sarnasemalt (neil on väiksem dispersioon ja standardhälve) kui metsikud hiired. Võimalik, et sarnasem reaktsioon on tingitud laborihiirte homogeensemast genofondist.

Standardhälbe ja dispersiooni omadusi: Olgu c konstant ja x uuritav tunnus. Siis

1. $s^2(cx) = c^2 s^2(x)$;
2. $s(cx) = cs(x)$;
3. $s^2(x + c) = s^2(x)$;
4. $s(x + c) = s(x)$;

Tähistuse $s^2(cx)$ all peame silmas suuruste cx dispersiooni (korrutame kõiki uuritava tunnuse x väärtuseid konstandiga c ja arvutame seejärel saadud suuruste dispersiooni), $s^2(x)$ on aga esialgsete x-tunnuse väärtuste dispersioon jne.

Teades vaid uuritava tunnuse keskväärtust (populatsiooni keskmist) ja standardhälvet, võime uuritava tunnuse väärtuste kohta öelda järgmist:

- vähemalt 3/4 uuritava tunnuse väärtustest asuvad keskväärtusele lähemal kui kaks standardhälvet (enamasti asub kahe standardhälbe kaugusel keskväärtusest umbes 95% vaatlustest);

- vähemalt 8/9 uuritava tunnuse väärtustest asub keskväärtusele lähemal kui kolm standardhälvet (enamasti asub kolme standardhälbe kaugusel keskväärtusest rohkem kui 99% vaatlustest).

2.4.3 Kvantiilid

α-kvantiiliks (α-*quantile*) nimetatakse sellist uuritava tunnuse väärtust, millest väiksemate väärtuste osakaal mõõtmistulemuste seas on α. Näiteks 0,1-kvantiil on selline uuritava tunnuse väärtus, millest väiksemad olid 10% meie mõõtmistulemustest ja 0,5-kvantiil on selline väärtus, millest väiksemaid väärtuseid on 50% (0,5-kvantiil on sama mis mediaan). Lisaks 0,5-kvantiilile kasutatakse sageli ka 0,25-kvantiili ja 0,75-kvantiili, mida kutsutakse ka **alumiseks ja ülemiseks kvartiiliks** (*quartile*).

Olukordades, kus tekib tahtmine kasutada (raporteerida) miinimumi ja maksimumi, soovitatakse kaaluda, kas poleks informatiivsem kasutada mõnda väikest ja suurt kvantiili, näiteks 0,05-kvantiili ja 0,95-kvantiili. Sel viisil on võimalik vältida mutantide ja andmesisestusvigade eksitavat mõju meid huvitava tunnuse kirjeldamisel ja langeb ära kiusatus andmete paremaks esitamiseks neid võltsida (ebamugavate vaatlus- või katsetulemuste "unustamise" teel).

Kuidas leida kvartiile? Üks traditsiooniline viis on järgmine: ülemise kvartiili hinnangu saame, kui leiame mediaanist suuremate vaatlustulemuste mediaani (variatsioonrea keskelt kuni lõpuni asuvad variatsioonrea elemendid), alumise kvartiiili saamiseks leiame mediaanist väiksemate vaatlustulemuste kvartiili. Kui mediaaniks (mediaani hinnanguks) osutub üks konkreetne variatsioonrea element, siis see vaatlus lisatakse kvartiilide arvutamisel nii mediaanist suuremate kui ka mediaanist väiksemate vaatluste sekka.

Kuna standardhälve ja dispersioon on (samuti nagu keskmine) tugevalt mõjutatavad üksikute vaatluste poolt, kasutatakse vahel alternatiivina ka kvartiiline vahet iseloomustamaks vaatluste hajuvust.

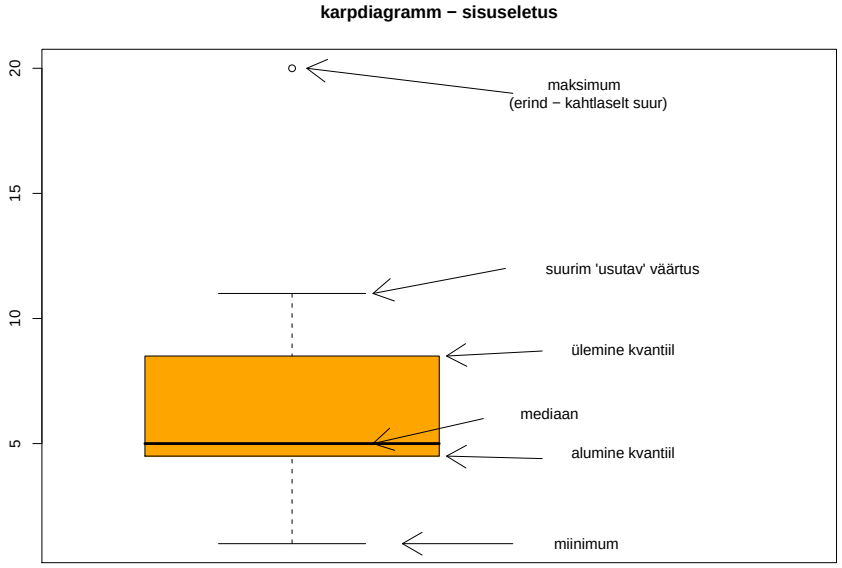
2.4.4 Karp-vurrud diagramm

Lisaks histogrammile kasutatakse pideva (vahel harva ka diskreetse) tunnuse jaotuse iseloomustamiseks ka **karp-vurrud** diagrammi (inglise k. *boxplot*). Karp moodustub ülemise ja alumise kvartiili vahele, karbi peale märgitakse

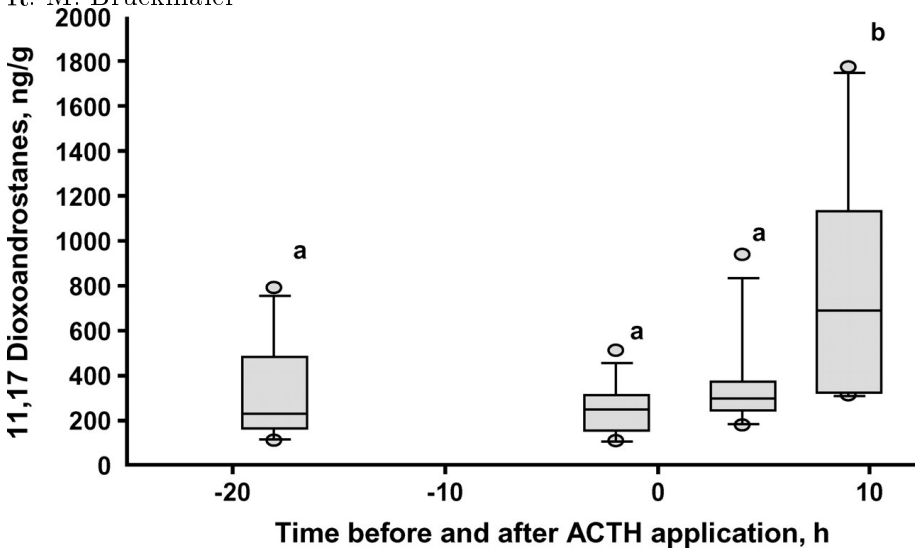
ka mediaani asukoht(alumine kvartiil oli väärtus, millest väiksemaid väärtuseid oli 25%, ülemine kvartiil oli aga väärtus, millest suuremaid väärtuseid oli 25%). Tekkinud karp "sisaldab" 50% vaatlustulemustest. Lisaks kantakse joonisele suurim ja väikseim vaatlustulemus, valimi miinimum ja maksimum. Miinimumi ja maksimumi ühendamisel karbiga tekivadki nn vurrud. Kui mõni valimis esinevatest uuritava tunnuse väärtustest on väga suur (või väga väike), siis ei tõmmata karpdiagrammi vurre mitte päris selle kahtlaselt suure väärtuseni, vaid mõne veidi pisema (paremini "usutava") väärtuseni. Sellisel juhul kantakse see üks (või enam) "kahtlaselt suurt" väärtust graafikule lihtsalt punktikestena. Kuidas arvuti otsustab, millal on vaatlus kahtlaselt suur (või kahtlaselt väike)? Ühtegi mõistlikku, sisuliselt põhjendatud meetodit sellise otsuse tegemiseks ei kasutata. Võib isegi öelda, et arvuti otsustab lihtsalt oma suva järgi (kuigi mõistagi mingit algoritmi kasutades).

Vaata jooniseid 2.5 ja 2.6.

Joonis 2.5: Karpdiagramm koos selgitustega



Joonis 2.6: Karpdiagrammi kasutusnäide artiklist J. Anim. Sci. 2004. 82:563-570 Coping capacity of dairy cows during the change from conventional to automatic milking D. Weiss*, S. Helmreich*, E. Möstldagger, A. Dzidic* and R. M. Bruckmaier*



Tabel 2.3: Tunnuse tüübile sobivad statistikud

	pidev	diskreetne	järjestustunnus	nominaalne
keskmine	+	+	-	-
mediaan	+	+	+	-
mood	+/-	+	+	+
dispersioon	+	+	-	-
standardhälve	+	+	-	-
kvartiilid	+	+	-	-
histogramm	+	+	-	-
tulpdiagramm	-	+	+	+
karpdiagramm	+	+	-	-

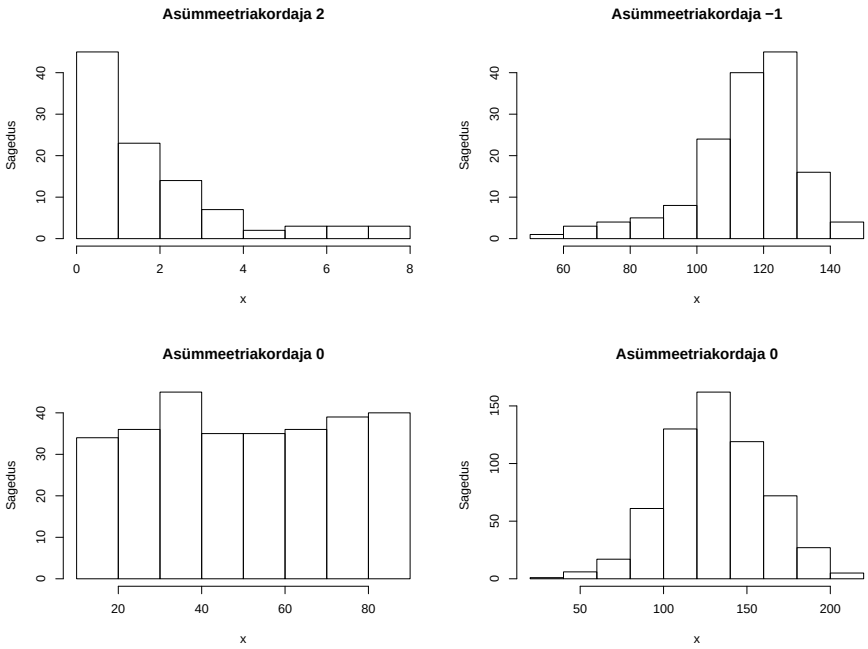
2.5 Teisi statistikuid

Asümmeetriakordaja a:

$$a = \frac{1}{(n-1)s^3} \sum_{i=1}^{n} (x_i - \bar{x})^3,$$

kus s on standardhälve. Sümmeetrilise jaotuse korral on asümmeetriakordaja a väärtus 0: $a = 0$. Kui esineb üksikuid väga suuri väärtuseid (tunnusel on raske saba paremal), siis on asümmeetriakordaja väärtus positiivne. Kui esineb üksikuid väga väikeseid mõõtmistulemusi, siis on asümeetriakordaja väärtus negatiivne. Vaata ka joonist 2.5. Kasutatakse, kui soovitakse rõhutada vaatluste jaotuse sümmeetrilisust/asümmeetrilisust.

Joonis 2.7: Asümmeetriakordaja väärtused erinevate jaotuste korral



Variatsioonikordaja c_v:

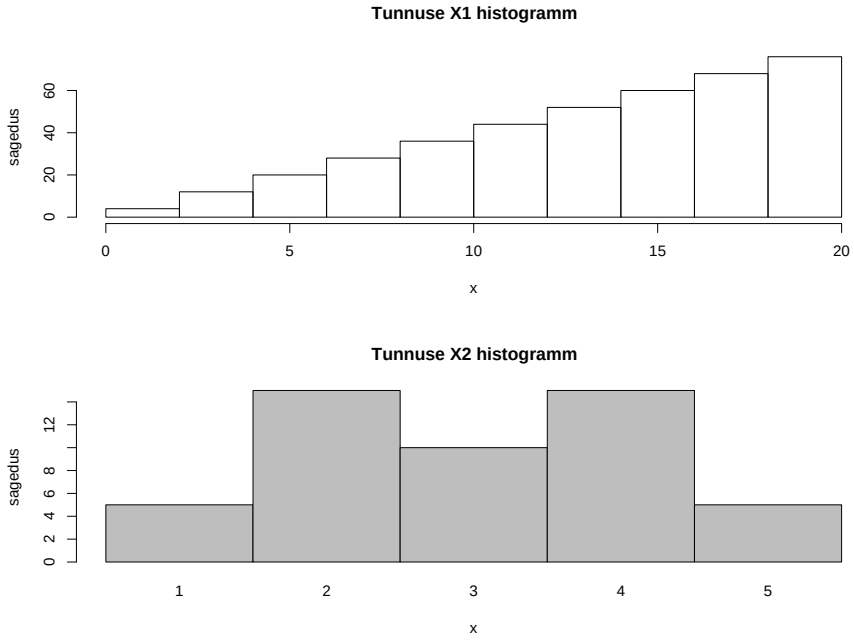
$$c_v = s/\bar{x}.$$

2.6 Ülesanded

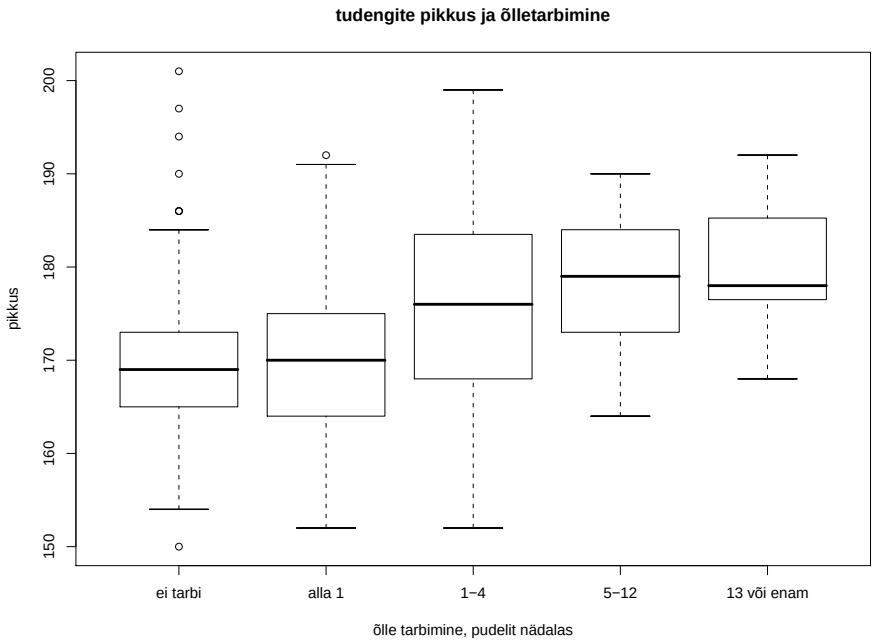
1. Vaata joonisel 1 antud histogramme. Leia nii $X1$ kui $X2$ oletuslik keskmine, dispersioon, mediaan, standardhälve.

2. Joonisel 2 on toodud karpdiagrammid tudengite pikkusele. Kuidas muutub tudengite pikkus sõltuvalt tarbitud õlle kogusest? Millest võiks see olla tingitud?

Joonis 2.8: Ülesanne 1 — milline võiks olla keskmine, standardhälve, mediaan ja dispersioon?



Joonis 2.9: Ülesanne 2. Tartu Ülikooli tudengid ja õlu.



Peatükk 3

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).

Esialgu ehk veidi üllatavalt selgub, et üks parimaid viise esindava valimi saamiseks on valimisse sattuvate objektide valimine juhuslikult.

Olgu igal populatsiooni kuuluval objektil võrdne võimalus sattuda valimi esimeseks objektiks. Olgu sõltumata sellest, kes või mis sattus valimi esimeseks objektiks (keda me esimesena mõõtsime) ikka igal uuritavasse populatsiooni kuuluval objektil ikka võrdne võimalus sattuda ka valimi teiseks objektiks (st. kui me esimesena küsitlesime Jaani, siis Jaanil on ikka teistega võrdne võimalus sattuda teiseks küsitletuks...), olgu kõigil samamoodi võrdne võimalus sattuda valimi kolmandaks objektiks jne. Sellisel viisil moodustatud valimit kutsutakse **lihtsaks juhuslikuks valimiks (tagasipanekuga)**, valimi moodustamise protseduuri tuntakse aga kui **lihtsat juhuslikku valikut**.

Enamik tavakasutuses olevaid statistikameetodeid eeldab, et uuritav valim on saadud lihtsa juhusliku valiku abil.

Kui kord valimisse sattunud objektil pole enam võimalik uuesti valimisse sattuda, kuid kõigil populatsiooni kuuluvatel isenditel on siiski võrdne valimisse valituks osutumise võimalus (ja kui ka mistahes populatsiooni kuuluval n objektil on samasuur võimalus koos samasse valimisse sattuda kui mistahes teisel n elemendilisel uuritavasse populatsiooni kuuluval isendite grupil)

siis räägitakse **tagasipanekuta lihtsast juhuslikust valimist** (näiteks kui inimesi valitakse uuringusse juhuslikult ja järgneva uuritava valimisel ei sõltu otsus sellest, kes eelnevalt valitud osutus, aga valikut tehakse siiski veel uurimata jäänute seast — ehk Jaani siiski kaks korda järjest ei küsitleta). Kui uuritav populatsioon on suur (näiteks lõpmatult suur), siis sobivad ka tagasipanekuta lihtsa juhusliku valimi uurimiseks needsamad statistilised meetodid, mis kõlbasid ka tagasipanekuga lihtsa juhusliku valimi uurimiseks. Kui uuritav populatsioon on väga väike, siis võib kaaluda spetsiaalseid lõpliku üldkogumi uurimiseks mõeldud statistiliste meetodite kasutamist.

Miks aitab lihtne juhuslik valik saada esindavat valimit?

Selle mõistmiseks tutvustame suurte arvude seaduseid.

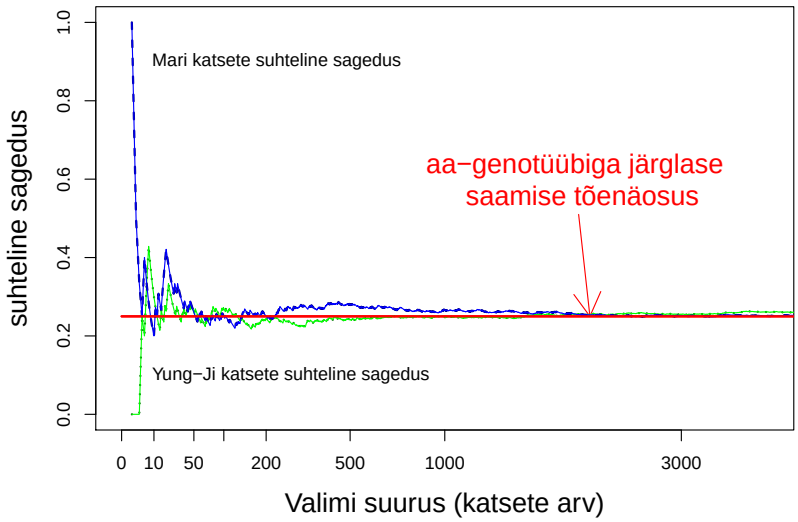
Oletame, et kordame mingit katset sõltumatult n korda ja loeme kokku, mitmel korral neist katsetest toimus meid huvitav sündmus A. Suurte arvude seadus (Bernoulli suurte arvude seadus) ütleb, et sündmuse A toimumise suhteline sagedus koondub katsete arvu lähenemisel lõpmatusele sündmuse A toimumise tõenäosuseks,

$$P(A) = \lim_{n\to\infty} \frac{\text{sündmuse A toimumiste arv}}{\text{katsete koguarv } n}.$$

Näiteks kui kaks teadlast, Mari Tartu Ülikoolist ja Yung-Ji Pekingi 23. riiklikust ülikoolist uurivad sama nähtust, näiteks heterosügootsetel vanematel (Aa ja Aa-genotüübiga vanematel) homosügootse aa-genotüübiga järglase saamise tõenäosust, siis mõlemal teadlasel heterosügootsete järglaste suhteline arvukus (aa-genotüübiga järglaste arv jagatud kõigi järglaste arvuga) koondub katsete arvu kasvades homosügootse järglase sündimise tõenäosuseks (mis antud juhul on 0,25), vaata ka joonist 3.1.

Suurte arvude seadusest järeldub, et kui korrates katset (nopime populatsioonist valimisse ühe uuritava objekti) palju kordi (sõltumatult — st me ei vali järgmisena nopitavat selle põhjal, kes meil varem valimis oli), siis koondub mistahes omaduse (loom on emane; katsetaim kasvab kolme kuuga pikemaks kui 25 cm; järglane on heterosügootne; ...) suhteline esinemissagedus võrdseks selle omaduse esinemistõenäosusega. Kui aga kõigil populatsiooni isenditel on võrdne võimalus sattuda valimisse, siis on sellesama omaduse esinemistõenäosus võrdne selle omadusega objektide osakaaluga populatsioonis.

Vaata ka alltoodud tabelit 3.1, kus on näha, kuidas lihtsa juhusliku valiku abil saadud valimis tunnuse jaotus muutub valimi suuruse kasvades üha sarnasemaks populatsiooni jaotusega (mida me üldjuhul ei tea).



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

Tabel 3.1: 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 4

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. See, kui mitu sordi XYZ tera läheb idanema meie valimis, huvitab meid vaid sedavõrd, kuivõrd ta aitab vastata üldisemale, populatsiooni puudutavale küsimusele.

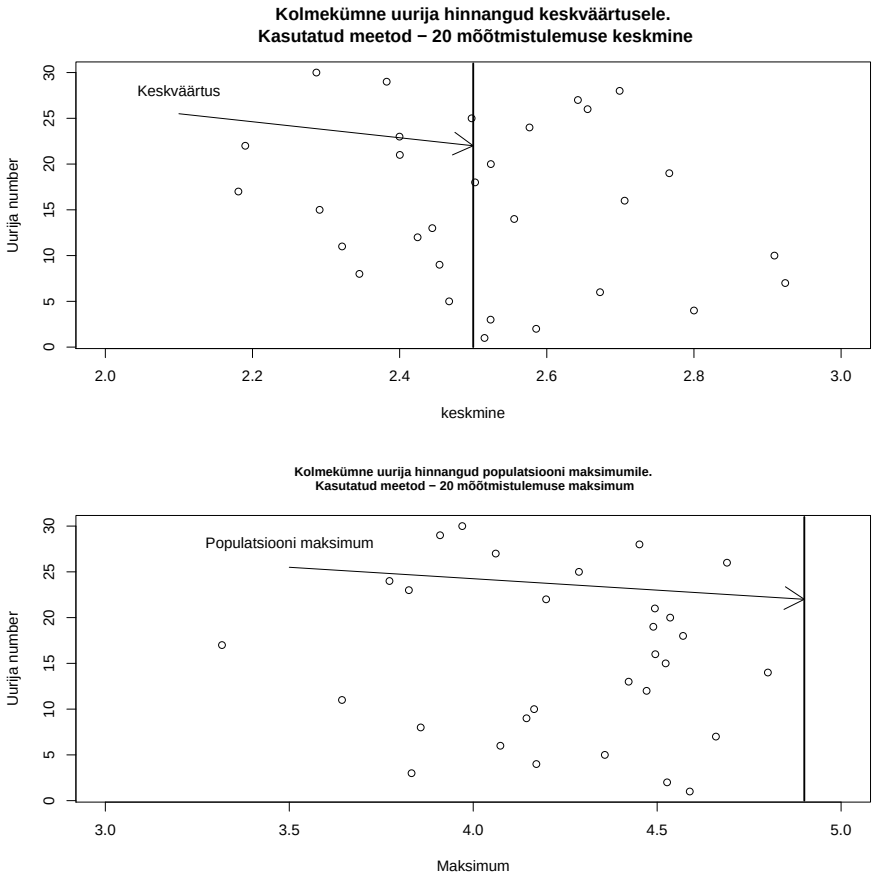
Populatsiooni parameetrite hindamiseks on mitmeid võimalusi. Tegelikku idanemisprotsenti võime hinnata kasutades kohvipaksu, valimi põhjal midagi arvutades või kasutades veel mõnda muud lähenemisviisi (näiteks alati pakkudes välja väärtust 3).

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. Võime vaadata, kui hästi erinevad hindamismeetodid suudavad õiget vastust ära arvata. Kergem on usaldada sellist metoodikat, mis teistega võrreldes 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).

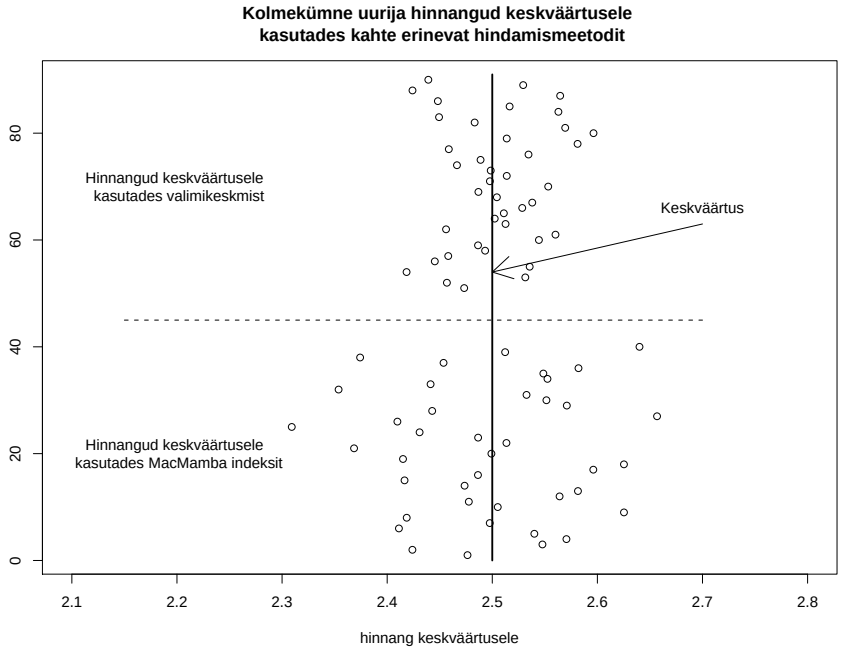
Joonis 4.1: Näide nihketa ja nihkega hinnangust



Kui räägime meid huvitavast (populatsiooni iseloomustavast) näitajast, siis on sageli tarvis teha vahet sellel, kas on tegemist tegeliku näitaja endaga või kõigest selle tegeliku näitaja (alati vigase) hinnanguga. Sagedamini kasutavate näitajate (statistikute) jaoks on isegi välja mõeldud erinevad tähistused. Näiteks populatsiooni keskmise ehk keskväärtuse tähistamiseks kasutatakse sageli sümboleid: $\mathrm{E}X$, μ; valimi keskväärtust tähistatakse aga sümboliga $\bar{x}$. Tabelis 4.1 on ära toodud tegelike arvnäitajate (populatsiooni parameetrite) enamlevinud tähistused ja nende valimi abil leitud hinnangute nimed ja tähistused.

Kui huvipakkuva väärtuse hinnanguks on üks konkreetne arv, näiteks kui

Joonis 4.2: Näide täpsemast ja vähemtäpsemast meetodist



Tabel 4.1: 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 $\hat{q}_\alpha$

hindame keskväärtust kasutades valimi keskmist, siis räägitakse, et tegemist on punkthinnanguga. Märkimaks, et tegemist on hinnanguga, kirjutatak-

se sageli arvkarakteristikut iseloomustava sümboli kohale laineke, katuseke, tärn vms.

4.1 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 4.1 *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 4.1.

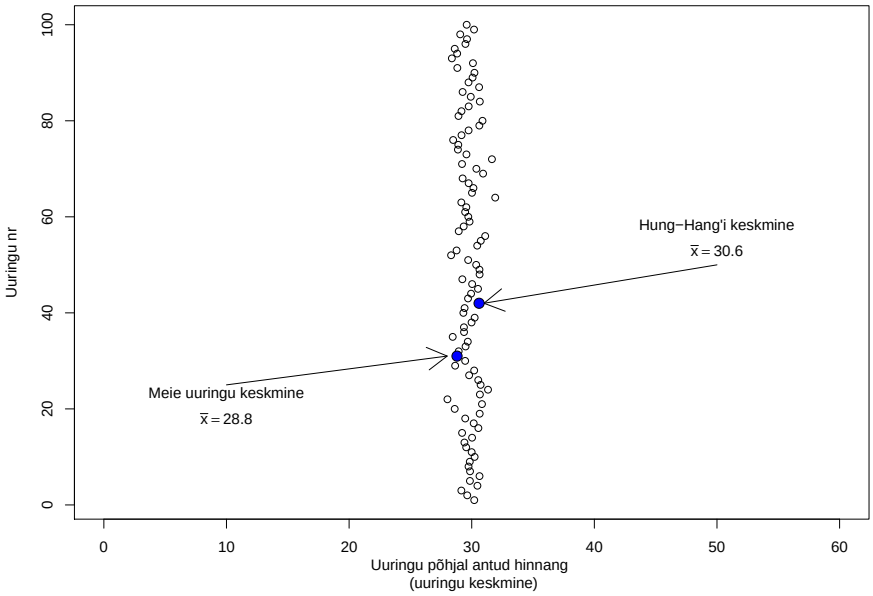
Teades kõigi nende uuringute tulemusi, on lihtne iseloomustada uuringu metoodika täpsust - kasutades näiteks uuringukeskmiste dispersiooni või standardhälvet, saame kirjeldada, kui kaugele võivad erinevad sama metoodikat kasutavate uuringute tulemused teineteisest tulla. Paraku pole meil uuringut teostades 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.

4.1.1 Dispersiooni omadusi

Populatsiooni dispersioon defineeritakse kui uuritava tunnuse väärtuste keskmine ruutkaugus keskväärtusest:

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

Joonis 4.3: Saja sama metoodikaga tehtud uuringu tulemused



Dispersiooni omadusi:

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:

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

4.2 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 5

Prognoosiintervall ja Usaldusintervall

5.1 Prognoosiintervall

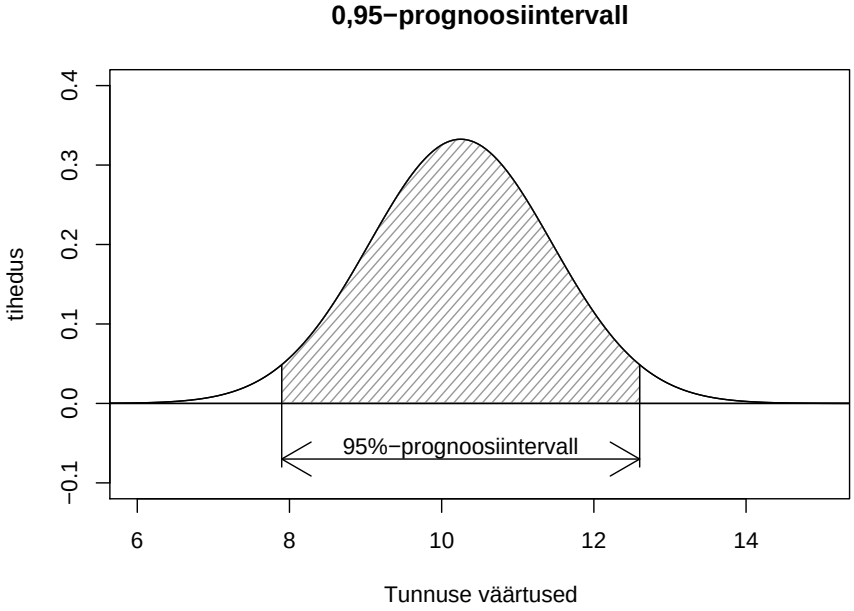
Unustame hetkeks populatsiooni parameetrite hindamise ja pöördume tagasi üksikvaatluste juurde.

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 5.1 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 5.1 välja lugeda soovitud (standardse normaaljaotuse) kvantiilid ning leida soovitud prognoosiintervalli. Märkus: standardse normaaljaotuse α-kvantiili tähistatakse traditsiooniliselt sümboliga z_α.

Joonis 5.1: 0,95-prognoosiintervall



Tabel 5.1: Standardse normaaljaotuse kvantiile 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:

$$\left(\mu + z_{\alpha/2}\sigma \ldots \mu + z_{1-\alpha/2}\sigma\right). \tag{5.1}$$

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 5.1, 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:

$$\left(\overline{X} + t_{\alpha/2;n-1}S\sqrt{1+\frac{1}{n}} \ldots \overline{X} + t_{1-\alpha/2;n-1}S\sqrt{1+\frac{1}{n}}\right). \tag{5.2}$$

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 5.1 on tõlgendatav kui vahemik, kuhu vahele jääb $(1-\alpha)\cdot 100\%$ vaatlustest, siis valemiga 5.2 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)!

5.2 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 5.2.

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 5.1 *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 5.2 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.

5.3 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 5.2: 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?

Peatükk 6

Hüpoteeside statistiline kontrollimine

Jälgides enda ümber toimuvat, elusloodust vaadeldes või aretusprotsessi juhtides, vahel ka teadusartikleid lugedes, tekib mõtlevatel inimestel paratamatult oletusi või hüpoteese ümbritseva maailma kohta. Oletus või kahtlus pole veel teadmine. Kuidas leida kinnitust oma kahtlustele? Kas kogutud andmed kinnitavad või hoopistükkis kummutavad meie oletuse/hüpoteesi?

6.1 Hüpoteeside kontrollimise filosoofiast

Kuidas saaks kontrollida, kas mingi oletus, teooria või hüpotees peab paika? Üks võimalus on järgmine. Oletatakse, et kontrollitav teooria peab paika. Arutluste abil leitakse, mis sellisel juhul (kui kontrollitav teooria peab paika) tulevikus juhtub, milliseid tulevaseid katsetulemusi me peaksime nägema, millise valimi peaksime saama, mida avastama tulevastel väljakaevamistel vms. Seejärel tehakse vajalik katse või vaatlus ning saadakse teada, kas teooria poolt ennustatud asi juhtus või mitte. Kui kontrollitava teooria ennustus ei pidanud paika, siis on teooria vale. Kui ennustus läks täkkesse, siis võib teooria kehtida (aga ei pruugi, sest paljud erinevad teooriad võivad viia sarnase ennustuseni). Seega kontrollitavat teooriat saab vaid kummutada, tavaliselt pole seda aga võimalik tõestada. Tõsi, kui kontrollitav teooria kannatab välja palju erinevaid kontrollimisi, siis tema usaldusväärsus tõuseb.

Hüpoteeside statistiline kontrollimine jälgib sarnast loogikat. Oletatakse, et mingi hüpotees (nimetagem seda nullhüpoteesiks) kehtib. Arutletakse, milliseid hinnangu väärtuseid (näiteks milliseid valimi keskmiseid) me nullhüpoteesi kehtides tõenäoliselt võiksime näha. Siis minnakse ja võetakse valim.

Kui valimi põhjal arvutatud hinnang tuli ootuspärane, selline nagu ta võiks tulla nullhüpoteesi kehtides, siis võib nullhüpotees õige olla. Kui aga meie hinnang polnud selline, nagu ta nullhüpoteesi kehtides oleks tõenäoliselt pidanud olema, siis lükkame nullhüpoteesi ümber. Teisisõnu — me usume siis, et kehtib väide, mis eitab nullhüpoteesi (alternatiivne hüpotees).

Hüpoteeside kontrollimist alustataksegi nullhüpoteesi ja alternatiivse hüpoteesi sõnastamisest. Nullhüpotees ja alternatiivne hüpotees peavad olema teineteist välistavad ja vähemalt üks neist peab kehtima. Traditsiooniliselt sõnastatakse hüpoteesid järgmiselt:

Nullhüpotees (H_0) - ka konservatiivne hüpotees. Väide, et praegu kehtiv traditsiooniline elu/tegu/mõtteviis tagab (vähemalt) sama hea tulemuse kui uuendajate/reformijate poolt pakutav lähenemisviis. Nullhüpoteesi kummutades — esivanemate traditsiooni hülgamist soovitades — võtame harilikult endale üsna suure vastutuse. Teadlasena muutuseid soovitades hakkame vastutama oma hea nimega.

Alternatiivne hüpotees (H_1) - kutsutud ka sisukaks hüpoteesiks. Väide, et uus lähenemisviis tagab parema tulemuse (on õigem) kui traditsiooniline, vaikimisi ja aruteludeta aktsepteeritav lähenemisviis. Üks noor ja ambitsioonikas teadlane arvatavasti unistab alternatiivse hüpoteesi tõestamisest — mis võiks olla veel ihaldusväärsem, kui näidata, et Sinu idee tagab parema tulemuse kui (teaduslike) esivanemate põlvkondadepikkune kogemus.

Näide 6.1 *Nullhüpotees: Mu vastaspartner mängib ausalt, ka ta täring on aus. Formaalselt kirja panduna H_0: P(sõber viskab täringuga 6 silma)=1/6.*

Alternatiivne hüpotees: Ta kasutab võltstäringuid! Formaalselt H_1: P(sõber viskab täringuga 6 silma)≠1/6.

Kuidas statistiliselt (vaid vaatlusandmetele tuginedes) saaks näites esitatud alternatiivset hüpoteesi tõestada? Siin ei ole tegelikult midagi tavamõistusele keerulist.

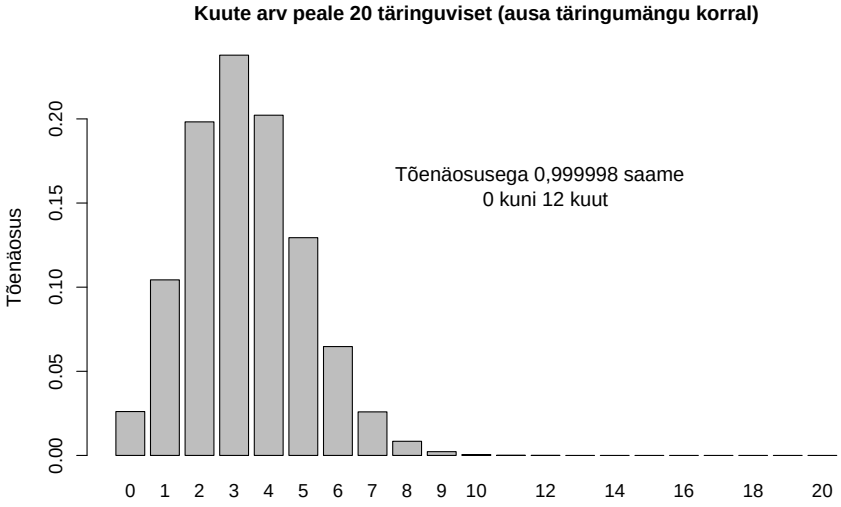
Näite 6.1 jätk...

Te mängite täringumängu. Mängupartner teeb heite ja saab kuus silma, teie kaotate. Halb õnn, mõtlete teie, ja mängite edasi. Vastasmängijal tuleb taas kuus, ja siis veel kord. Teil tekib kahtlus... Aga Te ei saa ju ometi niisama tõusta püsti ja esitada valjul häälel julma süüdistust... Teie nimi saaks määritud, kui kontrollimisel selguks, et täring on täiesti aus ja laua all pole peidus salakavalaid magneteid ega muid vidinaid. Alusetu süüdistus võiks igaveseks rikkuda Teie suhted selle auväärt ja kõrgel positsioonil inimesega kellega koos Te mängite. Te jätkate mängu, ja Teie kaaslane saab jälle kuue, ja jälle... Peale 20 täringuviset on vastasmängija saanud 19 korda tulemuseks kuus ja

vaid ühel korral on tema täring veeretanud kuuest väiksema numbri. Nüüd ei suuda te enam sellist enda pilkamist taluda — ka lauslollus määrib Teie nime.... Te järeldate, et Teie mängupartner teeb sohki ning Te süüdistate oma kaasmängijat pettuses.

Antud juhul töötab meie tunnetus jälgides sama loogikat, mida kasutab ka hüpoteeside statistiline kontrollimine. Juhul, kui täringumäng oleks aus, võiksime 20 viske jooksul näha ehk 9, äärmisel juhul kuni 12 kuut. Aga ausa mängu puhul kahekümne täringuviskega rohkem kuusi saada on peaaegu võimatu. Tõenäosusega 0,999998 jääb 20 ausa täringuviskega saadud kuute arv vahemikku [0...12], vaata ka joonist 6.1 (märkus: 20 ausa täringuviske jooksul saadud kuute arv on binoomjaotusega juhuslik suurus, $X \sim B(20{,}1/6)$). Seega, kui me näeme ennekuulmatut, lausa uskumatut tulemust — 19 korda visati kuus — siis me ei suuda enam ausasse mängu uskuda. Loomulikult pole ehk meie sisetunnetuse poolt tehtud tõenäosuse arvutused sedavõrd täpsed, aga küllap järeldus oleks ikkagi seesama — ausa mänguga siin tegemist pole.

Joonis 6.1: Kui mängukaaslane oleks aus...



Kirjeldatud lähenemisviisile on sisse ehitatud üks probleem — lugedes alternatiivse hüpoteesi tõestatuks, võime siiski kogemata teha vea. Ka ausa täringuga on põhimõtteliselt võimalik visata üheksateist korda järjest kuut. Loomulikult tuleks ausas teaduses iseloomustada, soovitavalt kvantitatiivselt, eksimise võimalikkust. See tingib vajaduse paari täiendava termini järgi.

Tehes oma otsust selle kohta, kas õige on nullhüpotees või alternatiivne hüpotees (kas mängupartner teeb sohki või mitte), võime eksida. Juhul, kui loeme alternatiivse hüpoteesi tõestatuks (ütleme, et sõber teeb sohki); aga tegelikult oli õige nullhüpotees (sõber tegelikult ei teinudki sohki) oleme teinud tõsise vea. Seda viga, alternatiivse hüpoteesi ekslikku õigekspidamist, kutsutakse *esimest liiki veaks* (*Type I error*).

Teist liiki viga (*Type II error*) tehakse siis, kui jäädakse ekslikult nullhüpoteesi juurde (ei süüdistata partnerit sohitegemises).

Tabel 6.1: Esimest ja teist liiki viga

	Tegelikult kehtib H_0	Tegelikult kehtib H_1
Jääme H_0 juurde	õige otsus	II liiki viga
tõestame H_1	I liiki viga	õige otsus

Enne hüpoteeside tõestamise juurde asumist tuleks otsustada, kui kindlad soovitakse oma tulemustes olla, kui maailmale minnakse alternatiivse hüpoteesi kehtimisest teatama. Matemaatilises keeles öeldult: fikseeritakse maksimaalne lubatud tõenäosus teha esimest liiki viga. Taolist lubatavat ülempiiri esimest liiki vea tegemise tõenäosusele kutsutakse olulisuse nivooks (*significance level*). Esimest liiki vea tegemise tõenäosus on maksimaalne muidugi siis, kui tegelikult kehtib nullhüpotees (kui alternatiivne hüpotees on õige, siis me ei saagi esimest liiki viga teha). Seega halvimas võimalikus situatsioonis — kui tegelikult on õige nullhüpotees — ei tohiks me alternatiivse hüpoteesi kasuks otsustada suurema tõenäosusega, kui valitud olulisuse nivoo lubab. Põhimõtteliselt võib muidugi kasutada väga erinevaid olulisuse nivoosid — perfektsionist võib kasutada näiteks olulisuse nivood 0,001 ja muidulahmija võib näiteks eelistada olulisuse nivood 0,25. Siiski on paljudes teadusvaldkondades/ teadusajakirjades esile kerkinud eelistatud olulisuse nivood — kõige sagedamini kasutatakse teaduskirjanduses olulisuse nivood 0,05.

Vähima olulisuse nivoo, mille korral me konkreetse eksperimendi korral oleksime veel saanud alternatiivse hüpoteesi tõestatuks lugeda, on *olulisustõenäosus* (*significance probability; p-value*). Kui olulisustõenäosus on väiksem kui valuläveks valitud olulisuse nivoo, võetakse vastu (loetakse tõestatuks) alternatiivne hüpotees.

Olulisustõenäosusele saab anda ka järgmise interpretatsiooni: olulisustõenäosus näitab, kui suur tõenäosus on näha meie poolt nähtud (või veel uskumatumat) tulemust siis, kui nullhüpotees kehtib. Pöördume hetkeks ta-

gasi täringumängu näite juurde. Ka aus mängija võib 20 täringuviske käigus saada üheksateist või enam korda tulemuseks kuus silma. See on lihtsalt väga ebatõenäoline — tõenäosus ausa mängu korral visata järjest 19 või enam korda järjest "6" on 0,0000000000000276. Seega võime öelda, et antud juhul on olulisustõenäosus $2{,}76 \times 10^{-14}$.

Statistiline test on seda parem, mida *tundlikum* ta on - mida väiksem on tõenäosus teha teist liiki viga, kui on fikseeritud maksimaalne lubatud tõenäosus teha esimest liiki viga. Statistilise testi võimsuseks kutsutakse tõenäosust lugeda tõestatuks alternatiivne hüpotees, kui tegelikult ongi õige alternatiivne hüpotees. Seega, mida suurema võimsusega (*power*) on test, seda parem ta on. Testi võimsus võib suuresti sõltuda sellest, milline on tegelikult uurijale huvipakkuv tundmatu protsess/väärtus. Näiteks, kui uuritakse keskkonnapoliitika mõju liigirikkusele, siis on mõistlikult ülesehitatud testide võimsus seda suurem, mida suurem on tegelikult poliitika mõju loodusele.

Näide: Mõõtmiste käigus koguti 50 istiku andmed ($n = 50$), uuritava tunnuse standardhälve olgu 1. Meid huvitavad hüpoteesid on järgmised (H_0: $\mu = 10$; H_1: $\mu \neq 10$). Tabelis 6.2 on antud I ja II liiki vea tegemise tõenäosused ja testi võimsus sõltuvalt populatsiooni tegelikust keskväärtusest.

Tabel 6.2: Esimest ja teist liiki vea tegemise tõenäoste sõltuvus uuritud tegelikkusest

Tegelik μ	9.95	10	10.05	10.1	10.2	10.3	10.5	11
I liiki vea tegemise tõenäosus (α)	-	0,05	-	-	-	-	-	-
II liiki vea tegemise tõenäosus	0,937	-	0,937	0,894	0,717	0,453	0,067	$> 0{,}001$
testi võimsus	0,063	-	0,063	0,106	0,283	0,547	0,933	0,999

Järgnevalt vaatleme mõningaid väga levinud statistilisi teste.

6.2 T-test hüpoteeside kontrollimiseks keskväärtuse kohta

Võib tulla ette olukordi, kus läheb tarvis kontrollida hüpoteese populatsiooni keskväärtuse kohta. Näiteks võib teaduskorüfee väita midagi aastaste istikute pikkuse keskväärtuse kohta, või on õpikus kirja pandud, milline peaks olema keskmine saagikus kirjeldatud kasvatusmeetodi korral. Kas poleks ahvatlev liig ülbeid autoriteete veidi õpetada ja nende väited kui valed kummutada?

Hüpoteesid keskväärtuse kohta saab kirja panna järgmiselt:

$$\begin{aligned} H_0: &\quad \mu = \mu_0 \\ H_1: &\quad \mu \neq \mu_0, \end{aligned}$$

kus μ_0 on mingi konkreetne väidetav number.

Juhul, kui nullhüpotees kehtib (tegelik keskväärtus ongi μ_0), ja vaatlusandmed on normaaljaotusega või valim on piisavalt suur ($n > 30$), siis on teisendatud juhuslik suurus (kutsugem teda t-statistikuks) $t := \frac{\bar{x}-\mu_0}{s}\sqrt{n}$ t-jaotusega juhuslik suurus (vaata usaldusintervalli kohta käivat peatükki):

$$t := \frac{\bar{x}-\mu_0}{s}\sqrt{n} \sim t_{n-1}.$$

Märkame, et t-statistiku väärtuse saame oma valimi põhjal kergesti välja rehkendada — teame ju nii oma valimi keskmist $\bar{x}$, valimi standardhälvet s kui ka valimi suurust n. Hüpoteesi sõnastades fikseerisime ka väärtuse μ_0-le.

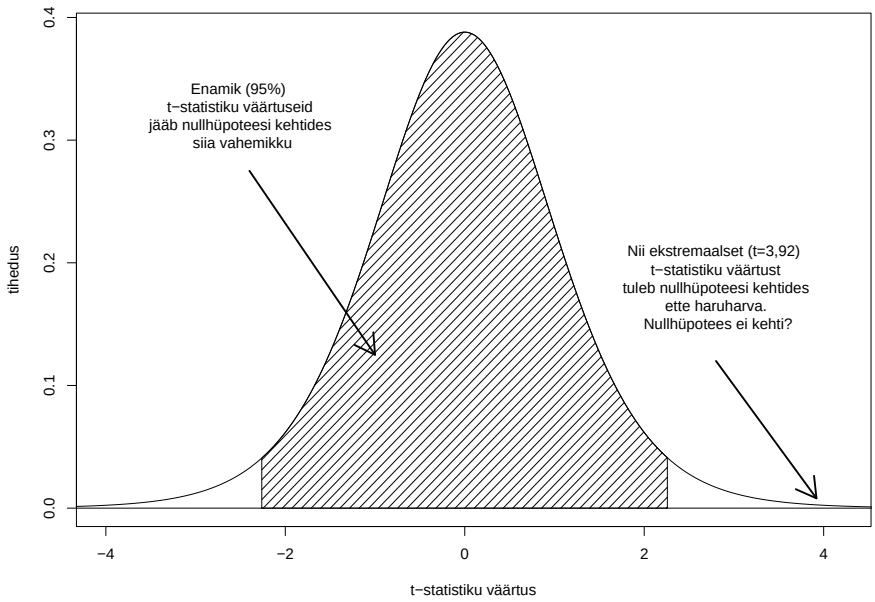
Kui nullhüpotees kehtib, siis iga uurija poolt (iga uue valimi põhjal arvutatud) leitud t-statistiku väärtus tuleb küll erinev, kuid enamikel juhtudel jäävad leitud väärtused üsnagi kitsastesse piiridesse, nulli lähedale. Valimi suuruse $n = 10$ korral on t-statistiku väärtuste jaotus (nullhüpoteesi kehtides) selline, nagu kujutatud joonisel 6.2. Kui meie valimi puhul leitud t-statistiku väärtus ei osutu tüüpiliseks / ootuspäraseks (näiteks t=3,92), siis on loomulik hakata kahtlema tehtud eelduse ($\mu = \mu_0$) paikapidavuses.

Mida pidada väga suureks või väga väikeseks, seda saab vaadata t-jaotuse tabelist. Kui nullhüpotees kehtib, siis jääb t-statistik t vahemikku

$$t_{\alpha/2;n-1} < t < t_{1-\alpha/2;n-1}$$

tõenäosusega $1-\alpha$. Väljapoole ülaltoodud vahemikku sattub t-statistik (nullhüpoteesi kehtides) haruharva, kõigest tõenäosusega α. Kui nüüd tõesti juhtub nii, et meie valimi põhjal leitud t on kas väiksem või suurem vaadeldavast

Joonis 6.2: t-statistiku väärtuste jaotus nullhüpoteesi kehtides (df=9)



kvantiilist ($t < t_{\alpha/2;n-1}$ või $t > t_{1-\alpha/2;n-1}$), siis tuleb tunnistada, et taolise situatsiooni tekkimise tõenäosus nullhüpoteesi kehtides on kaduvväike ja järelikult peab õige olema alternatiivne hüpotees. Seega kasutades olulisuse nivood α peaksime vastu võtma alternatiivse hüpoteesi.

Näide 6.2 *Sooviti uurida teatud kemikaali mõju närvisüsteemile. Korraldatud katses mõõdeti katseloomade (kasside) reaktsioonikiirust enne ja pärast uuritava kemikaali manustamist. Iga kassi jaoks leiti reaktsioonikiiruse muutus. Kui kemikaal närvisüsteemi ei mõjuta, siis peaks keskmine reaktsioonikiiruse muutus olema null ($\mu_0 = 0$):*

$$H_0: \quad \mu = 0$$
$$H_1: \quad \mu \neq 0.$$

Hüpoteeside kontrollimisel valime kasutatavaks olulisuse nivooks $\alpha = 0{,}05$.

Reaktsioonikiiruse muutused 15 katses osalenud kassi jaoks olid järgmised:

$$25, 27, -12, 8, 19, 4, -1, -4, 13, 12, 0.05, 2, -21, 24, -1.$$

Põhistatistikud vaadeldud valimi jaoks:

$$\begin{aligned} \bar{x} &= 6{,}33\ldots \\ s^2 &= 192{,}05 \\ s &= 13{,}86 \end{aligned}$$

Leiame t-statistiku väärtuse:

$$t = \frac{(6{,}33\ldots - 0)}{13{,}86} \times \sqrt{15} = 1{,}770\ldots.$$

Kas saame tõestatuks lugeda, et kemikaalil oli mõju kasside reaktsioonikiirusele? Valitud olulisuse nivoo (0,05) korral peame otsustamiseks tabelist üles otsima t-jaotuse 0,025- ja 0,975-kvantiilid:

$$\begin{aligned} t_{0{,}025;14} &= -2{,}14 \\ t_{0{,}975;14} &= 2{,}14. \end{aligned}$$

Kuna meie valimi põhjal leitud t-statistiku väärtus (1,77) jääb leitud kvantiilide vahele, siis järelikult ei saa me nullhüpoteesi ümber lükata (selline t-statistiku väärtus on täiesti mõeldav nullhüpoteesi kehtides). Järeldus: antud andmete põhjal ei saa tõestatuks lugeda, nagu mõjutaks vaadeldud kemikaal kasside reaktsioonikiirust.

Tähelepanek: Kui me oleksime valinud kasutatavaks olulisuse nivooks 0,1, siis oleksid võrdluses kasutatavate t-jaotuse kvantiilide väärtusteks tulnud -1,76 ja 1,76 ning me oleksime saanud alternatiivse hüpoteesi vastu võtta. Paraku peab kasutatav olulisuse nivoo olema määratud enne andmetega tutvumist ning olulisuse nivoo tagantjärgi tarkust kasutades on keelatud.

6.3 T-test sõltuvate valimite korral

Juba eelmises näites vaatlesime reaktsioonikiiruse muutuseid. Sarnased situatsioonid tulevad ette üllatavalt sageli — kas oleme teinud mõõtmisi enne ja pärast (millegi tegemist) ning soovime vaadata, kas uuritava tunnuse keskväärtus on muutunud; või oleme välja valinud sarnaste katseloomade paarid, ühte paarilist kasvatame ühtedes tingimustes, teist teistes. Soovime taas võrrelda, kas erinevates kasvutingimustes kasvatatud paariliste vahel on uuritava

tunnuse keskmises tasemes toimunud mingit muutust või mitte. Mida sarnasemaid mõõtmistulemusi ootaksime neil kahe paarilise mõõtmisel siis, kui "katsetingimuste muutmisel" mingit "mõju" poleks, seda kindlamalt suudame tõestada ka "mõju" olemasolu.

Sõltuvate valimite korral võib t-testi statistiku kirja panna järgmiselt:

$$t = \frac{\overline{x-y} - 0}{s(\overline{x-y})} = \frac{\overline{x-y} - 0}{s} \times \sqrt{n},$$

kus n on sarnaste paaride arv, x tähistab uuritava tunnuse väärtuseid paarilistel, kes viibisid ühtedes katsetingimustes ja y tähistab uuritava tunnuse väärtuseid paarilistel, kes viibisid teistes katsetingimustes. Tähistus $\overline{x-y}$ märgib paariliste katsetulemuste vahede keskmist. Kontrollitav hüpoteesipaar — kas uuritava tunnuse keskväärtus on sama, sõltumata katsetingimustest — on kirja pandav järgmiselt:

$$\begin{aligned} H_0: &\quad EX = EY \\ H_1: &\quad EX \neq EY. \end{aligned}$$

Ülaltoodud valemist arvutatud t-statistik on nullhüpoteesi kehtides (st. kui uuritava tunnuse keskväärtus mõlemas katsegrupis on sama) t-jaotusega, vabadusastmete arvuga n-1 (vaatluspaaride arv - 1).

6.4 T-test sõltumatute valimite keskväärtuste võrdlemiseks

Alati pole võimalik moodustada sarnast katsealuste paare selliselt, et uuritava "mõju" puudumisel annaksid mõlemad katsealused sama (või peaaegu sama) katsetulemuse. Näiteks võib meid huvitada, kas sordi A saagikuse keskväärtus on samasuur kui sordi B saagikuse keskväärtus. Aga sarnaste põllulappide/põldude valik võib osutuda keeuliseks. Või on katse juba toimunud — ühte sorti kasvatati 6, teist 23-l põllul. Kuidas siis kontrollida hüpoteese keskväärtuste võrdsuse kohta:

$$\begin{aligned} H_0: &\quad EX = EY \\ H_1: &\quad EX \neq EY? \end{aligned}$$

Sellisel juhul kasutatakse t-testi sõltumatute vaatluste jaoks. Statistilise testi konstrueerimiseks on kaks võimalust. Ühel juhul eeldatakse, et mõlemas grupis uuritava tunnuse hajuvus on samasuur, teisel juhul taolist täiendavat eeldust ei tehta.

6.4.1 T-test sõltumatute vaatluste jaoks, võrdne hajuvus.

Eeldades, et uuritava tunnuse hajuvus mõlemas vaadeldavas grupis (mõlemas populatsioonis) on samasuur, võib püstitatud hüpoteesipaari kontrollimiseks kasutada järgmist teststatistikut:

$$t = \frac{\bar{x} - \bar{y}}{s(\bar{x} - \bar{y})},$$

kus

$$s(\bar{x}-\bar{y}) = \sqrt{s^2(\bar{x} - \bar{y})} = \sqrt{s^2(\bar{x}) + s^2(\bar{y})} = \sqrt{s^2/n_1 + s^2/n_2} = s/\sqrt{\frac{n_1 + n_2}{n_1 n_2}}.$$

Viimases valemis tähistab s ühist hinnangut standardhälbele — st hinnang on leitud mõlemat valimit kasutades. Kahe valimi ühine hinnang standardhälbele on leitav järgmise valemi abil:

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2}{n_1 + n_2 - 2}}.$$

Juhul, kui kehtib nullhüpotees, siis on taolisel viisil arvutatud t-statistiku jaotuseks t-jaotus, vabadusastmete arvuga $df = n_1 + n_2 - 2$. Kui arvutatud

t-statistiku väärtus on väga ekstremaalne (ebausutavalt suur/väike võrreldes t-jaotuse poolt lubatud kõrvalekaldega), siis järeldatakse, et tehtud eeldused olid väärad — st. rühmade keskväärtused on erinevad.

Sellisel viisil tehtud t-testi tulemus on usaldusväärne, kui: a) uuritava tunnuse hajuvus mõlemas rühmas on ligikaudu samasuur (mõlema sordi saagikuste varieeruvus on sarnane) b) uuritav tunnus (saagikus) on kas normaaljaotusega või on uuringus kasutatud palju katsepõlde ($n_1 + n_2 > 30$).

6.4.2 T-test sõltumatute vaatluste jaoks, hajuvus erinevates populatsioonides võib olla erinev (Waldi test).

Kahe populatsiooni keskväärtuste kontrollimiseks on võimalik konstrueerida t-testi ka ilma võrdse hajuvuse eeldust tegemata. Kasutatav t-statistik näeks sellisel juhul välja järgmine:

$$t = \frac{\bar{x} - \bar{y}}{s(\bar{x} - \bar{y})} = \frac{\bar{x} - \bar{y}}{\sqrt{s^2(\bar{x}) + s^2(\bar{y})}} = \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2/n_1 + s_y^2/n_2}}.$$

Asja komplitseerib mõnevõrra tõdemus, et leitud statistik pole nullhüpoteesi kehtides enam täpselt t-jaotusega juhuslik suurus, ta on kõigest peaaegu t-jaotusega. Millist vabadusastmete arvu kasutada? Leidub erinevaid soovitusi (mis võivad anda veidi erinevaid tulemusi), üks levinumaid on nn Satterthwaite'i meetod. See meetod ütleb, et testimisel tuleks kasutada järgmise valemi abil leitud vabadusastmete arvu:

$$df = \frac{(s^2(\bar{x}) + s^2(\bar{y}))^2}{\{s^2(\bar{x})\}^2/(n_1 - 1) + \{s^2(\bar{y})\}^2/(n_2 - 1)}.$$

Enamasti teostab rehkendused arvuti ja sestap pole valemi keerukus eriliseks takistuseks. Kui arvutatud t-statistiku väärtus on väga ekstremaalne (ebausutavalt suur/väike võrreldes t-jaotuse poolt lubatud kõrvalekaldega), siis järeldatakse, et tehtud eeldused olid väärad - st. rühmade keskväärtused on erinevad. Sellisel viisil tehtud t-testi tulemus on usaldusväärne, kui: b) uuritav tunnus (näiteks saagikus) on kas normaaljaotusega või on uuringus osalenud inimesi palju.

6.4.3 Sobiva T-testi valimine

Kas vaatlused on paari pandavad (tehtud samal inimesel, tehtud kaksikutel või sama pesakonna kahel järglasel, tehtud sarnaste põllulappide paaridel)? Kui jah, siis kasuta t-testi sõltuvate valimite jaoks (inglise k. paired t-test). Kui tegemist pole vaatlustega samal isendil või sarnaste isendite paaridel,

siis otsusta, kas uuritava tunnuse hajuvus võiks mõlemas võrreldavas rühmas (populatsioonis) olla sama? Kui jah, siis kasuta T-testi sõltumatute võrdse hajuvusega vaatluste jaoks. Kui ei, siis kasuta T-testi, mida kirjeldatud alalõigus 6.4.2.

6.5 Teisi teste

Wilcoxoni (Mann-Whitney) test — t-testi analoog, aga ei eelda, et uuritav tunnus oleks normaaljaotusega (aga eeldab, et uuritav tunnus on pidev). Juhul kui Wilcoxoni (Mann-Whitney) test otsustab alternatiivse hüpoteesi kasuks — ühes populatsioonis on uuritava tunnuse väärtused suuremad kui teises — võib osutuda raskeks kirjeldada, kuidas (ja kui suurelt) need kaks populatsiooni siis ikkagi erinevad teineteisest. Erinevus Wilcoxoni testi mõttes ei pruugi tähendada keskväärtuste erinevust, samuti ei saa sellest, et Wilcoxoni test lükkas ümber nullhüpoteesi järeldada veel mediaanide erinevust. Seega statistiliselt olulise testitulemuse korrektne interpreteerimine võib osutuda keerukaks.

Kolmogorov-Smirnovi test — kontrollib hüpoteese, kas uuritava tunnuse jaotus kahes populatsioonis on sama või mitte. Seega, kui ühes populatsioonis on näiteks uuritava tunnuse hajuvus suurem kui teises, siis Kolmogorov-Smirnovi test avastab (suure valimi korral) erinevuse. Või kui populatsioonide keskväärtused on erinevad. Või kui leidub mõni muu erinevus uuritava tunnuste jaotustes. Kolmogorov-Smirnovi testi kasutatakse vahel ka kontrollimaks, kas meie uuritava tunnuse jaotus võiks olla mõni kindel, hästituntud teoreetiline jaotus. Testi eeldused: Nõuab, et uuritav tunnus oleks pidev. Paljud statistikaprogrammid võivad anda kahtlaseid tulemusi, kui esineb (palju) kokkulangevaid uuritava tunnuse väärtuseid. Märkus: Kolmogorov-Smirnovi test on enamasti üsna madala võimsusega — tema võime märgata erinevusi on kehvavõitu. Sestap vajame alternatiivse hüpoteesi tõestamiseks suurt valimit.

F-test — kontrollib hüpoteese uuritava tunnuse hajuvuse kohta. Näiteks võime F-testi abil kontrollida, kas uuritava tunnuse hajuvus kahes populatsioonis on samasuur või mitte (näiteks kas kahe sordi saagikused — näiteks üle aastate — on sama stabiilsed või on üks vaadeldav sort tundlikum keskkonnamõjudele kui teine). Eeldused — uuritava tunnuse jaotuseks normaaljaotus.

6.6 Mitmese võrdluse probleem

Tõestanud ühe või teise hüpoteesi oleme harjunud tõestatud väidet õigeks pidama sõltumata sellest, kui mitu korda me oleme varem üritanud meid huvitavat hüpoteesi tõestada. Samuti ei oska me harilikult tähelepanu pöörata sellele, milliseid ja kui paljusid teisi hüpoteese me oleme proovinud tõestada. Paraku on hüpoteeside statistilise kontrollimise korral olukord keerulisem, sest kasutatav metoodika annab eriti jäärapäisele uurijale "järgi" ja võib mõnigi kord lugeda alternatiivset hüpoteesi tõestatuks ka siis, kui tegelikult kehtib nullhüpotees.

Näide 6.3 *Kujutame ette olukorda, kus kehtib nullhüpotees (imeväetisest pole kasu). Alternatiivse hüpoteesi kehtivuses veendunud inimene korraldab uuringu ja jõuab tulemuseni, et peab jääma nullhüpoteesi juurde. Oma tõe õigsusesse uskudes korraldab ta selle peale uue uuringu ja veel ühe uuringu jne. Korraldades 20 korrektselt läbiviidud uuringut, igaüks neist teostatud olulisuse nivool 0,05, on tõenäosus, et ükski neist kahekümnest uuringust ei otsusta alternatiivse hüpoteesi tõestatuks lugeda (1-0,05)20=0,358. Kui asjast huvitatud jäärapäine teadlane otsustab avaldada vaid sellise uuringu tulemused, mis alternatiivse hüpoteesi õigeks loevad, siis teeb see hüpoteetiline jäärapäine teadlane esimest liiki vea tõenäosusega 1-0,358=0,642.*

Näide 6.4 *Otsitakse põhjuseid, mis võiksid soodustada haiguse tekkimist. Viidi läbi uuring, kus haigetel ja tervetel taimedel mõõdeti sadade erinevate tunnuste väärtused (kui toitainerikkas mullas nad kasvavad, kui pikk on neid kasvatav talumees, palju sadas vihma eelmisel kuul jne). Andmed kokku kogutud, hakatakse t-testi abil võrdlema, kas haigeks jäävad taimed kasvasid toitainerikkamas mullas kui terved, kas haigeid taimi kasvatas pikem talumees kui terveid jne. Oletame, et andmestikus oli 100 tunnust, mis kirjeldasid taimede kasvutingimusi. Iga kasvutingimuse korral võrreldakse terveid ja haigeid ja võrdlemisel kasutatakse olulisuse nivood 0,05. Oletame, et ükski uuritud kasvutingimus tegelikult ei mõjuta (ei muuda) võimalust haigeks jääda (terveid taimi kasvatanud talunike pikkuse keskväärtus oli sama mis haigeid taimi kasvatanud talunike pikkuse keskväärtus jne). Vaatamata sellele võiksime antud uuringu puhul tõestatuks lugeda umbes 5 haiguse tekkimist soodustavat tegurit (tõenäosus, et saame tõestada vähemalt ühe küsimuse korral keskväärtuste erinevuse ehk teeme esimest liiki vea on 0,994).*

Üks levinud meetod mitmese võrdluse ohte vältida on Bonferroni meetod. Juhul kui on tarvis kontrollida n erinevat hüpoteesi ja soovime, et tõenäosus

teha üht või enamat valeotsust ohtlikus suunas (I-liiki vea tegemise tõenäosus) ei ületaks α, siis tuleb Bonferroni meetodi järgi teha iga üksiktest kasutades olulisuse nivood α/n.

Tasub tähele panna, et mida rohkem küsimusi (üksikteste) esitame, seda väiksemaks läheb üksiktesti teostamisel tarvitatav olulisuse nivoo. Seega läheb ka raskemaks alternatiivse hüpoteesi vastuvõtmine. Seetõttu tuleks tehtavate testide arvu hoida siiski võimalikult väike, kontrollides statistilise hüpoteeside kontrollimise protseduuri abil vaid neid hüpoteese, mis teoreetiliste arutelude põhjal võiksid kõige enam huvi pakkuda. Vältida tuleks sihipäratut katsetamist ja proovimist, sest siis võib statistilise hüpoteeside testimise protseduur anda halbu tulemusi.

Bonferroni meetod kipub paljudes olukordades olema veidi liiga konservatiivne, st. tema kasutamisel on maksimaalne esimest liiki vea tegemise tõenäosus enamasti väiksem kui α (5%). Alternatiivina Bonferroni meetodile võib kasutada näiteks Bonferroni-Holmi meetodit. Bonferroni-Holmi meetod eeldab, et kõigepealt teostatakse kõik testid ja leitakse nende olulisustõenäosused. Saadud olulisustõenäosused järjestatakse kasvavalt: $p_1 \leq p_2 \leq .... \leq p_k$. Otsused nullhüpoteesi kasuks või kahjuks tehakse siis kasutades olulisuse nivoosid

$$\alpha/k, \alpha/(k-1), \alpha/(k-2), ..., \alpha/2, \alpha.$$

Summarne esimest liiki vea tegemise tõenäosus on ka Bonferroni-Holmi meetodi puhul maksimaalselt α. Bonferroni-Holmi meetod on aga tsipa vähem konservatiivne kui Bonferroni meetod, st. Bonferroni-Holmi meetod otsustab praktikas sagedamini alternatiivse hüpoteesi kasuks kui Bonferroni meetod. Miinuseks on asjaolu, et Bonferroni-Holmi meetod eeldab, et kõik testid on enne tema kasutamist tehtud. See välistab tema kasutamise olukorras, kus soovime teha osa teste tänaste andmete pealt ja osad testid kahe aasta pärast (siis kui on rohkem andmeid kogunenud).

Kui teostavate testide arv kasvab, väheneb Bonferroni meetodi rakendamisel kiiresti ka kasutatav olulisuse nivoo ja alternatiivse hüpoteesi tõestamine osutub sageli äärmiselt raskeks (nõuab tohutult paljude vaatluste olemasolu). Selle tõttu pole statistilised meetodid mitte eriti sobivad katse/eksituse meetodil teaduse tegemiseks (proovime, kas midagi õnnestub). Ideaaljuhul jõutakse arusaamisele hüpoteeside kehtivuse kohta kasutades mittestatistilisi argumente, lähtudes näiteks teoreetilise bioloogia tõekspidamistest, ja statistilist hüpoteeside kontrolli rakendatakse pigem võimalike eksimuste tuvastamiseks oma arutluskäigus.

Praktikas kipub olukord olema muidugi vastupidine, vähesed viitsivad mõelda ja suhteliselt rohkem on neid, kellele meeldib niisama proovida ja katsetada. Sestap on ka olemas tugev surve arutu proovimise jaoks sobivate statistikameetodite järgi. Mitmese testimise kontekstis on üheks huvitavaks kontseptsiooniks nn "valeavastuste määra" piiramine. (False Discovery Rate). Idee selle termini taga on järgmine: las juhtugu esimest liiki vigasid, see pole suur probleem, aga teeme nii, et vastuvõetud alternatiivsete hüpoteeside seas poleks ekslikult vastuvõetud alternatiivseid hüpoteese mitte rohkem kui mingi lubatud protsent (5%). Ehk teisisõnu öeldes me kontrollime valepositiivsete ehk ekslike "avastuste" osakaalu kõigi "avastuste" seas. Antud teemaga seotud statistilised meetodid on praegu kiirelt arenev valdkond ja vähesed statistikaprogrammid pakuvad võimalust mitmese testimise probleemi ohjamiseks piirata valepositiivsete osakaalu ehk limiteerida False Discovery Rate-i. Üks lihtne võimalus antud lähenemist ise kasutada oleks järgmine:

1. järjesta testide poolt raporteeritavad olulisustõenäosused kasvavalt,

$$p_1 \leq p_2 \leq .... \leq p_k.$$

2. Kirjuta välja kriitilised suurused $a/k, 2*a/k, 3*a/k, ..., a$, kus a näitab maksimaalset nn False Discovery Rate´i.

3. Võrdle esimest, teist jne järjestatud olulisustõenäosust vastava kriitilise suurusega kuni jõuad paarini, kus olulisustõenäosus on suurem kui vastav kriitiline suurus. Loe kõigi talle eelnenud hüpoteeside jaoks alternatiivne hüpotees tõestatuks ja tema ning järgnevate testide puhul jää nullhüpoteesi juurde.

Näide 6.5 *Teostati 10 testi, nende olulisustõenäosused (järjestatult) on järgmised: 0,0002; 0,008; 0,01; 0,04; 0,045; 0,12; 0,22; 0,25; 0,62; 0,83.*

Vastavad kriitilised väärtused on (a = 0,05 korral): 0,005; 0,01; 0,015; 0,02; 0,025; 0,03; 0,035; 0,04; 0,045; 0,05

Esimene olulisuse nivoo, mis on suurem vastavast kriitilisest väärtusest, on 4. olulisuse nivoo. Seega testide 1-3 korral võtame vastu alternatiivse hüpoteesi ja testide 4-10 jaoks jääme nullhüpoteesi juurde.

Vahel on võimalik kontrollitavad hüpoteesid selliselt ümber sõnastada, et tekiks üks uus kontrollitav hüpotees ja paljude üksiktestide asemel tuleb teha üksainus test. Üks selline erijuht on keskmiste mitmene võrdlus. Oletame, et meil on m gruppi, mille keskmist taset tahame võrrelda. Näiteks õpetavad m õpetajat sama õppeainet m-grupile tudengitele. Soovime teada, kas kõigi õpetajate õpilased saavad peale kursuse läbimist (keskmiselt) võrdselt hästi

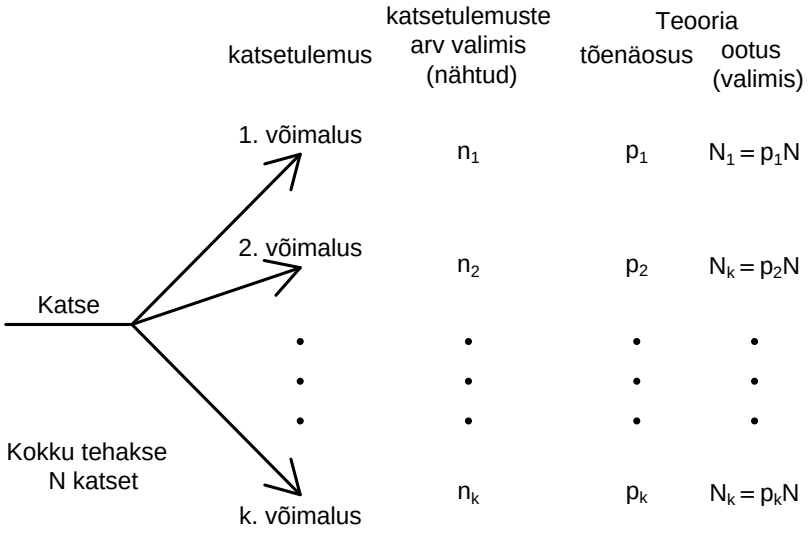
hakkama, või leidub mõni õpetaja, kes teistest paremini oma ainet seletab. Üks võimalus soovitu väljaselgitamiseks on teha $m(m-1)/2$ t-testi võrdlemaks kõigi õpetajate töö tulemusi kõigi teiste õpetajate töö tulemustega. Sealjuures tuleks kasutada mõnda mitmese võrdluse meetodit (näiteks Bonferroni meetodit). Teine võimalus oleks kasutada t-testi edasiarendust dispersioonanalüüsi üheainsa hüpoteesipaari kontrollimiseks. Dispersioonanalüüsi vaatleme lähemalt ühes järgnevas loengus.

6.7 Hii-ruut test

Üks universaalsemaid ja sagedamini kasutust leidev test on hii-ruut (χ^2-test, inglise keeles ka *chi-square test*).

Oletame, et sooritataval katsel on k erinevat võimalikku tulemust (lilleseemnest kasvab kas valge, punane või roosa õis — $k = 3$; sündiv linnupoeg on kas isane või emane — $k = 2$ jne). Mõnikord on võimalik teooriat kasutades leida, milline peaks olema ühe- või teise katsetulemuse tulemise tõenäosus. Sellisel juhul on võimalik hii-ruut testi abil kontrollida, kas vaatlustulemused on kooskõlas teooria ennustustega või mitte — kas erinevus teooria ja valimis nähtu vahel võiks olla tingitud valimi juhuslikkusest või on erinevus liiga suur. Vaata ka joonist 6.7.

Joonis 6.3: Tegelikud katsetulemused ja teooria ennustus



Vaatame mõningaid näiteid hüpoteesidest, mida saab testida kasutades hii-ruut testi.

Näide 6.6 *Grupi teadlaste arvates on lõvilõua õie värv määratud ühe geeni poolt. Sellel geenil on kaks alleeli, tähistame neid a ja A. Juhul, kui taime genotüüp on AA, peaks tal olema punane õis, genotüübiga Aa lill peaks olema roosa õiega ja aa-genotüübiga lill võiks olla valge. Selle hüpoteesi kontrollimiseks ristasid teadlased roosade õitega (heterosügootseid) taimi. Kui nende*

oletus peab paika, peaks järglaste jaotus olema kooskõlas Mendeli seadustega, vt tabel 6.3.

Tabel 6.3: Mendeli seaduste paikapidavuse kontrollimine

	AA (punane õis)	25% järglastest
Aa x Aa	Aa (roosa õis)	50% järglastest
	aa (valge õis)	25% järglastest

Näide 6.7 *Soovitakse kontrollida, kas uuritav populatsioon on Hardy-Weinbergi tasakaalus (antud geeni suhtes alleelidega a ja A). Tähistame tõenäosust, et populatsioonist juhuslikult valitud geenialleel on A tähega p. Juhul, kui populatsioon oleks Hardy Weinbergi tasakaalus, peaks genotüüpide esinemistõenäosused olema sellised, nagu antud tabelis 6.4.*

Tabel 6.4: Hardy-Weinbergi tasakaalu kontrollimine

genotüüp	esinemistõenäosus
AA	p^2
Aa	$2p(1-p)$
aa	$(1-p)^2$

Märkus: Tõenäosust p saame hinnata oma valimi põhjal — $\hat{p} = \frac{2\#\{AA\}+\#\{Aa\}}{2N}$.

Kuidas siis hii-ruut test kontrollib sedaliiki hüpoteese? Esmalt leiame hii-ruut statistiku väärtuse,

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - N_i)^2}{N_i}$$

Juhul, kui teooria peab paika, siis terooria poolt ennustatud juhtude arv N_i peaks olema ligilähedaselt õige ja vahed $n_i - N_i$ tulevad väikesed ning ka hii-ruut statistiku χ^2 väärtus tuleb väike. Seevastu juhul, kui teooria ei pea paika, siis kipuvad erinevused nähtu ja oodatu vahel olema (absoluutväärtuselt) suured, ning hii-ruut statistiku väärtus tuleb suur.

Soovides kasutada antud statistikut testimaks teooria paikapidavust, peame selgitama, kui suur peab teststatistiku väärtus olema selleks, et me enam ei

tohiks uskuda teooria paikapidavusse. Selle selgitamiseks tuleb esmalt määrata veel üks vajalik parameeter, mida kutsutakse vabadusastmete arvuks. Veidi lihtsustatud (kuid peaaegu alati korrektselt töötav) eeskiri hii-ruut testi vabadusastmete arvu leidmiseks on järgmine:

$$\begin{aligned} df &= \text{Erinevate võimalike tulemuste arv } (k) \\ &\quad -\text{valimi põhjal hinnatud parameetrite arv suuruste } N_i \text{ leidmiseks} \end{aligned}$$

Saamaks näite 6.3 jaoks kontrollitava teooria poolt ennustatavat juhtude arvu, peame oma andmetest piiluma ühte numbrit — valimi suurust. Teades valimi suurust N, saame teooria põhjal öelda, mitu punase õiega, mitu roosa õiega ja mitu valge õiega lille Mendeli seaduste järgi peaks valimis olema. Seega antud näite korral on vabadusastmete arv

$$\begin{aligned} df &= 3 \text{ (kolme värvi õitega järglasi võib esineda)} \\ &\quad -1 \text{ (valimi suuruse hindamine)} \\ &= 2 \end{aligned}$$

Näite 6.4 korral oleme sunnitud sageduste N_i leidmiseks hindama (kasutades valimit) kahte parameetrit: valimi suurust N ja alleeli A esinemissagedust p:

$$\begin{aligned} df &= 3 \text{ (3 erinevat genotüüpi)} \\ &\quad -2 \text{ (valimi suurus ja alleeli A esinemissagedus)} \\ &= 1 \end{aligned}$$

Lisaks peame testi läbiviimiseks fikseerima olulisusenivoo — määratlema, kui kindlad me tulemustes tahame olla, enne kui julgeme nullhüpoteesi kummutada. Teades usaldusnivood, vabadusastmete arvu ja Hii-ruut statistiku väärtust saame otsuse teha kasutades hii-ruut jaotuse tabelit. Kriitilised väärtused on toodud tabelis 6.5.

Näide 6.8 *Ühel euroopas esineval liblikaliigil (Panaxia dominula) esineb ühe geeni mutatsioon, mis muudab liblika tiibade mustrit. Alleelidega AA liblikail on tiibadel suured valged täpid. Genotüübiga aa isenditel täpid puuduvad ja nende asemel on tiivad ühtlaselt tumedad. Genotüübiga Aa isenditel täpid esinevad, kuigi on väiksemad kui AA tüüpi isenditel. Uurijaid huvitab,*

Tabel 6.5: χ^2 - statistiku kriitilised väärtused h

Vab.-astmeid (df)	$P(X > h) = 0{,}05$	$P(X > h) = 0{,}01$
1	3,841	6,635
2	5,991	9,210
3	7,815	11,345
4	9,488	13,277
5	11,070	15,068
6	12,592	16,812
7	14,067	18,475
8	15,507	20,090
9	16,919	21,666
10	18,307	23,209
12	21,026	26,217
14	23,685	29,141
16	26,296	32,000
18	28,869	34,805
20	31,410	37,566
25	37,652	45,624
30	43,773	50,892
35	49,802	57,342
40	55,758	63,691
45	61,656	69,957
50	67,505	76,154
60	79,082	88,379
70	90,531	100,425
100	124,32	135,807

kas üks või teine tiivamuster annab liblikale mingit evolutsioonilist eelist või on liblikate populatsioon Hardy-Weinbergi tasakaalus. Selle selgitamiseks läks uurija Ford 1971. aastal aasale ja määras 1612 liblika genotüübid:

Vastamaks meid huvitavale küsimusele tuleb leida geeni A esinemistõenäosus:

$$p := p(A) = (2 * 1469 + 138)/(1612 * 2) = 0{,}954$$

Nüüd saame leida näites 2 toodud valemeid kasutades, kui palju me ühe või teise genotüübiga isendeid oleksime oodanud olevat oma valimis, kui

Tabel 6.6: härra Fordi andmed

AA-tüüpi isendeid:	1469
Aa-tüüpi isendeid:	138
aa-tüüpi isendeid:	5

Hardy-Weinbergi tasakaal oleks kehtinud. Genotüübiga AA isendite oodatav proportsioon oleks $p_i = p^2 = 0{,}954^2 \approx 0{,}9103$*, ootuspärane genotüübiga AA isendite arv valimis oleks* $N_i = Np_i = 0{,}91031612 \approx 1467{,}4$*, erinevus tegeliku ja oodatava vahel on* $n_i - N_i = 1469 - 1467{,}4 = 1{,}6$*. Samamoodi võime jätkata arvutusi ka teiste genotüüpide tarvis, vaata ka tabel 6.7.*

Tabel 6.7: Härra Fordi arvutused

	tegelik n_i	oodatav proportsioon	oodatav arv N_i	erinevus $n_i - N_i$
AA-tüüpi isendeid	1469	0,9103	1467,4	1,6
Aa-tüüpi isendeid	138	0,0876	141,2	-3,2
aa-tüüpi isendeid	5	0,0021	3,4	1,6

Nüüd saab välja arvutada ka hii-ruut statistiku väärtuse:

$$\chi^2 = \frac{1{,}6^2}{1467{,}4} + \frac{(-3{,}2)^2}{141{,}2} + \frac{1{,}6^2}{3{,}4} = 0{,}827.$$

Olles valinud usaldusnivooks 0,05 (tüüpiline teadusartiklites kasutatav usaldusnivoo), võrdleme oma saadud hii-ruut statistikut tabelis antud väärtustega. Kuna antud juhul on vabadusastmete arv 1 (3-2=1), siis peame võrdlema leitud statistiku väärtust (0,827) kriitilise väärtusega, mis on antud reas df=1 (3,841). Kuna 3,841>0,827, siis järeldame, et erinevus teooria poolt ennustatava ja vaatlustulemuste vahel on väike. Me pole suutnud nullhüpoteesi ümber lükata, Hardy-Weinbergi tasakaal võib kehtida. Seega ei saa nende vaatlustulemuste põhjal öelda, et ühe tiivamustriga järglastel oleks evolutsiooniline eelis teistsorti tiivamustriga liblikate üle.

Antud klassikalise näite andmed pärinevad raamatust E. B. Ford (1971) Ecological genetics. Chapman and Hall, London.

6.7.1 Hii-ruut testi eeldused

Hii-ruut test baseerub asümptootikal, st. selleks et testi tulemus oleks korrektne, peab valim olema suur. Kui suur peaks olema valim, et praktikas saaks kasutada hii-ruut testi? Harilikult soovitatakse, et igat võimalikku väärtust esineks H_0 kehtides ootuspäraselt enam kui viiel korral ($N_i \geq 5$ iga i korral). Mida teha, kui oletades kontrollitava teooria kehtimist peaks antud valimi suuruse juures mingit väärtust esinema vähem kui 5 korda? Üks võimalus (lisaks valimi suurendamisele) oleks kombineerida kokku mõned harvemad väärtused üheks uueks väärtusklassiks. Seejärel tuleks kontrollida, kas kombineeritud väärtuste esinemissagedus vastab teooria poolt ennustatavale. Loomulikult peame võimalike väärtuste arvu vähendades vähendama ka kasutatavat vabadusastmete arvu. Lisaks mainitud eeldusele peab loomulikult olema tegemist nõuetekohaselt leitud (juhusliku) valimiga.

Järeldus: viimases näites toodud arvutus ei pruugi olla (päris) korrektne, sest genotüübiga aa isendeid ootasime valimis olevat vaid 3,4 tükki (mida on alla 5).

6.7.2 Hii-ruut test seose olemasolu kontrollimiseks

Vahel soovitakse testida, kas kahe (nominaalse, järjestus-) tunnuse vahel eksisteerib statistiline seos või mitte (kas ühe tunnuse väärtuse teadmine aitab öelda midagi selle kohta, milline võiks olla teise tunnuse jaotus — ehk teisisõnu — kas ühe tunnuse väärtuse muutudes muutub teise tunnuse jaotus või mitte). Näiteks võime soovida kontrollida, kas alamliigiti erineb saakloomade eelistus (kas erinevatel alamliikidel on erinev saakloomade jaotus) või võime testida, kas eksisteerib seos inimese töölemineku viisi (jala/rattaga/bussiga/autoga) ja tema tervisliku seisundi (suurepärane; hea; keskmine; halb; väga kehv) vahel.

Kontrollimaks hii-ruut testiga seose olemasolu tunnuste X ja Y vahel (olgu nende tunnuste võimalikud väärtused tähistatud vastavalt sümbolitega $1,2,\ldots,k$ ja $1,2,\ldots,l$) peame leidma, milline on mistahes väärtuste komplekti $(X = i, Y = j)$ saamise tõenäosus siis, kui kontrollitav hüpotees (väide: tunnused on sõltumatud) kehtiks. Juhul, kui tunnused X ja Y oleks tõepoolest sõltumatud, siis peaks tunnuse X jaotus olema alati samasugune, ükskõik, milline siis ka tunnuse Y väärtus ka pole. Juhul, kui seost tunnuste vahel pole, on tõenäosus näha tunnuse X väärtust i üks ja seesama sõltumata sellest, milline on tunnuse Y väärtus:

$$P(X = i|Y = 1) = P(X = i|Y = 2) = \ldots = P(X = i|Y = l) \;\; (= P(X = i)).$$

Ülaltoodud valemis tähistab $P(X = i|Y = j)$ nn tinglikku tõenäosust — juhul kui teame, et tunnuse Y väärtus on j, siis tõenäosus, et tunnuse X väärtus tuleb i on $P(X = i|Y = j)$.

Kui X ja Y on sõltumatud, siis $P(X = i|Y = j) = P(X = i)$ ja tõenäosus, et populatsioonist juhuslikult valitud isendi korral saame sellise looma/taime/inimese, kelle puhul $X = i$ ja $Y = j$ on leitav järgmiselt:

$$P(X = i, Y = j) = P(X = i|Y = j)P(Y = j) \overset{sõltumatud}{=} P(X = i)P(Y = j).$$

Seega saame leida tunnuste mistahes väärtuste puhul, millise tõenäosusega üht- või teistsugune väärtuste komplekt peaks esinema nullhüpoteesi (seost tunnuste vahel pole) paikapidamisel. Ja edasi jätkame juba nii, nagu hii-ruut testi tehakse – leiame oodatud arvud ($N_{X=i,Y=j} = NP(X = i)P(Y = j)$), leiame erinevused teooria poolt ennustatud sageduste ja tegelikult valimis nähtud sageduste vahel ning arvutame hii-ruut statistiku väärtuse. Vaatame vaid üle vabadusastmete arvu leidmise - mitut numbrit või parameetrit me peame enne oma valimi põhjal leidma, enne kui saame arvutada suurused $N_{X=i,Y=j}$? Esiteks peame teadma oma valimi suurust N. Teiseks peame hindama tunnuse X jaotuse ehk tõenäosused $P(X = 1), \ldots, P(X = k)$. Paneme aga tähele, et viimast tõenäosust $P(X = k)$ me ei pea piiluma oma valimist — kuna $P(X = 1) + \ldots + P(X = k) = 1$ siis $P(X = k) = 1 - P(X = 1) + \ldots + P(X = k - 1)$ ja seega peame valimi põhjal hindama kõigest $k - 1$ tõenäosust (ja viimase saame juba eelnevate põhjal leida). Sama juhtub tõenäosuste $P(Y = 1), \ldots, P(Y = l)$ leidmisel — viimase tõenäosuse saame leida kasutades teisi. Seega valimi põhjal tuleb leida kokku 1(valimi suurus N) + $k - 1$(tunnuse X jaotus: $P(X = 1), \ldots, P(X = k - 1)$) + $l - 1$(tunnuse Y jaotus: $P(Y = 1), \ldots, P(Y = l - 1)$) = $k + l - 1$ parameetrit ja antud juhul tuleb hii-ruut testi vabadusastmete arvuks

$$\begin{aligned} df &= kl - (k + l - 1) \\ &= kl - k - l + 1 \\ &= (k - 1)(l - 1). \end{aligned}$$

Näide 6.9 *Vaatame, kas saame tõestada, et tudengite alkoholitarbimise ja nende soo vahel eksisteerib seos. Algandmed on esitatud tabelis 6.8.*

Esmalt hindame nii õlletarbimise kui ka soolise jaotuse (tabelid 6.9 ja 6.10).

Nüüd leiame tõenäosused $P(sugu = i, õlu = j)$. Leitud tõenäosused on toodud tabelis 6.11.

Nüüd leiame teooria (sõltumatus) poolt ennustatavad tudengite arvud $N_{sugu=i,õlu=j}$, vaata tabel 6.12.

Tabel 6.8: Seos tudengite soo ja nädalase õlletarbimise vahel

sugu | õlu	ei tarbi	alla pudeli	1-4	5 või enam	kokku
naine	241	222	43	6	512
mees	25	44	49	31	149
kokku	266	266	92	37	661

Tabel 6.9: Õlletarbimise jaotus

x	ei tarbi	alla pudeli	1-4	5 või enam
P(õlu=x)	0.4024 =266/661	0.4024	0.1392	0.0560

Tabel 6.10: Sooline jaotus

x	naine	mees
$P(sugu = x)$	0.7746=512/661	0.2254

Tabel 6.11: Oodatavad tõenäosused sõltumatuse korral

sugu|õlu	ei tarbi	alla pudeli	1-4	5 või enam	kokku
naine	0.3117=0.4024*0.7746	0.3117	0.1078	0.0434	0.7746
mees	0.0907=0.4024*0.2254	0.0907	0.0314	0.0126	0.2254
kokku	0.4024	0.4024	0.1392	0.0560	1

Tabel 6.12: Sõltumatuse korral oodatavad tudengite arvud

sugu|õlu	ei tarbi	alla pudeli	1-4	5 või enam	kokku
naine	0.3117*661=206.04	206.04	71.26	28.66	512
mees	59.96	59.96	20.74	8.34	149
kokku	266	266	92	37	661

Ja lõpuks võime leida hii-ruut statistiku väärtuse:

$$\chi^2 = \frac{(241 - 206.04)^2}{206.04} + \frac{(25 - 59.96)^2}{59.96} + \ldots + \frac{(31 - 8.34)^2}{8.34} = 161.$$

Vabadusastmeid oli $df = (2 - 1) * (4 - 1) = 3$*, hii-ruut statistiku kriitiline väärtus on 7,815, meie statistiku väärtus on aga palju suurem — palju suurem kui sõltumatute tunnuste puhul oleks võinud tulla. Järeldus: nullhüpotees ei saa kehtida, tunnused ei saa olla sõltumatud. Tunnuste sugu ja õlletarbimine vahel esineb statistiline seos. Seega saame väita, et eri soost tudengite õlletarbimisharjumused on erinevad.*

6.8 Ülesanded

1. Ristati haplotüüpidega Aa, Bb ja Aa, Bb isendeid. Saadud andmed on toodud tabelis 6.13. On teada, et alleelid A ja a päranduvad vastavalt mendeli seadustele, st ristamisel Aa x Aa saadud järglane on tõenäosusega 0,25 genotüübiga AA, tõenäosusega 0,50 genotüübiga Aa ja tõenäosusega 0,25 genotüübiga aa. Sama reegli järgi päranduvad ka alleelid B ja b. Küsimus: kas geeni A alleelid (A,a) ja geeni B alleelid (B, b) päranduvad sõltumatult või on tegemist nn geeniaheldusega (linkage)?

Tabel 6.13: Vaatlusandmed

B|A	AA	Aa	aa	kokku
BB	23	7	2	32
Bb	10	32	7	49
bb	3	10	20	33
kokku	36	49	29	114

Kuidas muutub teststatistik ja vabadusastmete arv, kui me ei eelda üksikute alleelide pärandumist vastavalt mendeli seadustele? Näiteks võib olla võimalik, et genotüüpi AA esineb mingil põhjusel heterosügootsete vanemate lastel "liiga" palju? Ehk mis juhtub siis, kui teeme tavalise hii-ruut testi kahe muutuja sõltumatuse kontrollimiseks? (Vihje — kirjeldatud kahe juhu korral tulevad nii teststatistiku väärtused kui ka vabadusastmete arvud erinevad, samuti võime jõuda erinevate otsusteni!)

Peatükk 7

Seosed tunnuste vahel

Peatükk 8

Lihtne lineaarne regressioon

Siin peatükis õpime käejoonte järgi pikka või lühikest elu ennustama
ehk
regressioonmudel, mudeli olulisuse testimine ja tema eeldused, determinatsiooni- ja korrelatsioonikordaja

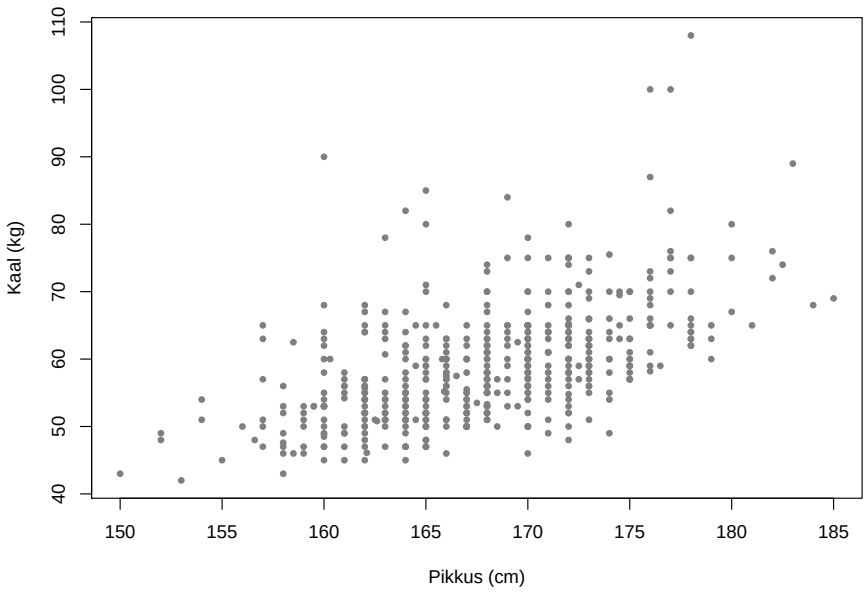
Nähes inimest, kelle nägu kaunistavad rohked kortsud, oskame üht-teist tema kohta arvata ka siis, kui me midagi muud temast ei tea. Võime arvata, et ta on vanem, elukogenud ja küllap ka veidi konservatiivne (arvatavasti eelistab ta ikka veel paberraamatut lugeda ning ehk ta ei ole kuulnudki iPad'ist või Kindlest...). Mõistes ühe tunnuse (kortsude arv näos) seost teise tunnusega (vanus) on meil võimalik "ennustada" meile tundmatu tunnuse väärtuseid (oletada teise inimese vanust). Selline ennustamine võiks olla eriti meelepärane siis, kui ühe tunnuse mõõtmine on suhteliselt lihtne ja odav (piisab ühest pilgust inimese näkku märkamaks ta kortse), samas kui teise tunnuse mõõtmine on raske, kallis või ohtlik (eks proovige küsida daamilt tema vanust!)

Vahel on erinevate tunnuste vahelised seosed tõepoolest intuitiivselt mõistetavad ja piisab ka nende seoste põhjal tehtud ligikaudsetest järeldustest. Mõnikord aga vajame võimalikult täpseid tulemusi (hästi, on elukogenum, aga kui palju elukogenum?), või on tunnustevaheline seos meie jaoks üldse mõistatuseks — kas see üldse eksisteerib, ja kui, siis milline see on? Sellistel juhtudel — kui täpsus on oluline või kui intuitsioon ei aita — võib endale appi paluda regressioonanalüüsi.

8.1 Sissejuhtaus

Kui uurimise all on kaks pidevat tunnust, mille vahelist seost tahame mõista, on enamasti tark esmalt vaadata tunnustevahelist hajuvusgraafikut. Joonisel 8.1 on toodud Tartu Ülikoolis õppivate naistudengite pikkuse ja kaalu vahelist seost kirjeldav hajuvusgraafik (*scatterplot*).

Joonis 8.1: Hajuvusgraafik



Hajuvusgraafikult näeme, et pikemad tudengid kipuvad ka kaalukamad olema. Seost nende kahe tunnuse vahel võiks justnagu kirjeldada sirge abil. Sageli ongi seost kahe tunnuse vahel üsna hästi võimalik kirjeldada sirge abil (muidugi mitte alati, aga esimeseks lähendiks sobib sirge üllatavalt sageli).

Igaüks võiks tõmmata seost iseloomustava sirge läbi hajuvusgraafikul kujutatud punktipilve. Üldises vestluses, sõprade seltsis, võime tõepoolest rahulduda sellise käega veetud sirgega. Paraku tõmbab üks inimene veidi ühtemoodi sirge kui teine, isegi läbi sellesama punktipilve. Tekivad küsimused — milline sirge on parem, milline joonis viiks täpsemate prognoosideni?

Parem on muidugi see sirge, mis viib võimaldab uusi, tulevasi väärtuseid, võimalikult täpselt prognoosida. Paraku see, milline sirgetest osutus paremaks, selgub alles tagantjärgi, siis kui prognoositavad sündused on juba

toimunud. Enamasti tahaksime teha valiku sirgete vahel enne seda hetke.

Üks võimalus valida kõikvõimalike seost iseloomustavate sirgete vahel on järgmine. Vaatame, milline sirge prognoosib kõige paremini olemasolevaid andmeid. Peaksime muidugi täpsustama, mis mõttes täpsemalt (kuidas mõõdame täpsust). Üheks parimaks täpsuse mõõdupuuks on osutunud vähimruutude nõue — eelistame sirget, mille puhul prognoosivigade ruutude summa oleks minimaalne (alternatiivne sõnastus: keskmine ruutviga oleks minimaalne). Täpsemalt, kui i. vaatlus on y_i, tema prognoos regressioonisirge abil (kui meid prognoosimisel abistava tunnuse väärtus on x_i) on $\hat{y}_i = \hat{c}_0 + \hat{c}_1 x_i$, siis valime $\hat{c}_0$ ja $\hat{c}_1$ väärtused selliselt, et prognoosivigade

$$\begin{aligned} e_i &:= y_i - \hat{y}_i \\ &= y_i - \hat{c}_0 + \hat{c}_1 x_i \end{aligned}$$

ruutude summa

$$\sum_{i=1}^{n} e_i^2$$

oleks minimaalne. Joonisel 8.1 kujutaud pikkuse ja kaalu andmetega sobib kõige paremini järgmine regressioonsirge:

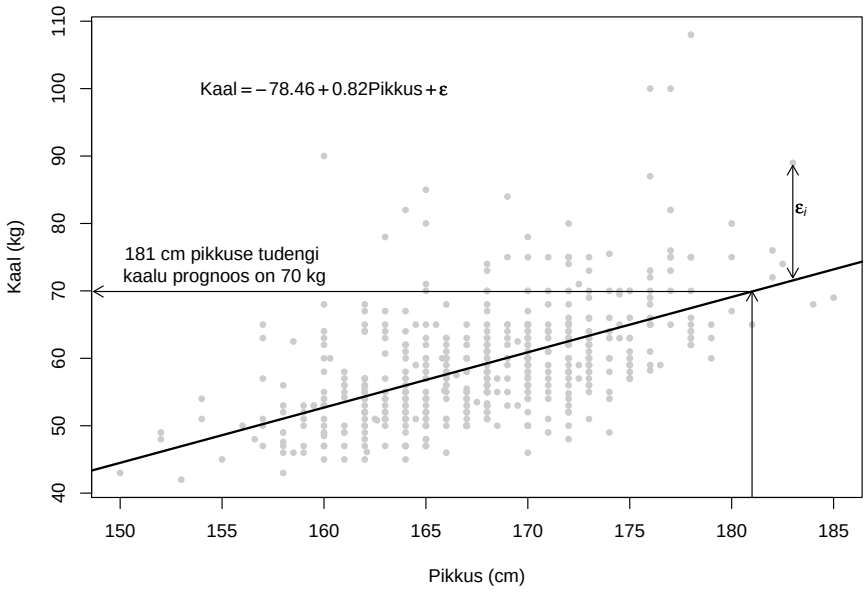
$$\hat{kaal} = -78{,}46 + 0{,}82 \cdot pikkus,$$

ehk sirge, kus vabaliige $c_0 = -78{,}46$ ja sirge tõus $c_1 = 0{,}82$, vaata ka joonist 8.10.

8.2 Regressioonanalüüsi mudel

Kui kahe pideva tunnuse vahel eksisteerib statistiline seos, siis peaks ühe tunnuse väärtuste muutudes muutuma ka teise tunnuse väärtuste jaotus. Sageli (kuid mitte alati) seisneb muutus selles, et ühe tunnuse väärtuste kasvades kipuvad teise tunnuse väärtused olema ka suuremad (või vastupidi väiksemad). Näiteks pikemad inimesed kipuvad ka rohkem kaaluma; suuremates metsades kipub kasvama rohkem puid; mida rohkem inimene joob karastusjooke, seda lühem kipub olema tema eluiga. Kuigi suuremas metsas kipub olema rohkem puid kui väiksemas metsas, ei tähenda see veel seda, et ei võiks esineda mõnda suurt ja hõredat metsa, kus kasvab puid vähem kui mõnes tillukeses tihedalt väikseid noori puid täis metsas.

Joonis 8.2: Prognoosivigade ruutude summat minimiseeriv sirge



Regressioonanalüüsi mudel kirjeldabki, kuidas muutub Y-tunnuse keskväärtus siis, kui vaatleme erineva X-tunnuse väärtustega objekte. Lihtsa lineaarse regressioonanalüüsi mudel ütleb, et Y-tunnuse keskväärtus muutub lineaarselt X-tunnuse väärtuste muutumisel:

$$\mathrm{E}Y = c_0 + c_1 x. \tag{8.1}$$

Kui räägime naistudengite pikkusest ja kaalust, siis võiks sobiv mudel välja näha järgmine:

$$\mathrm{E}kaal = -78{,}46 + 0{,}82 \cdot pikkus. \tag{8.2}$$

Kuidas tõlgendada antud mudelit? Kui vaataksime vaid 150cm pikkuseid tudengeid, siis selliste tudengite keskmine kaal on antud mudeli järgi

$$-78{,}46 + 0{,}82 \cdot 150 = 44{,}54kg.$$

See ei tähenda, et iga 150cm pikkune tudeng peaks kaaluma täpselt 44,54 kg. Antud mudel lihtsalt väidab, et 150cm pikkuste tudengite keskväärtus on 44,54kg. Samuti väidab ta midagi ka 151cm pikkuste tudengite keskmise

kaalu kohta. Sentimeetri võrra pikemate tudengite keskmine kaal peaks antud mudeli järgi olema $-78{,}46 + 0{,}82 \cdot 151 = 44{,}36kg$ ehk $0{,}82kg$ rohkem.

Toodud näidet võime üldistada andmaks regressioonsirge tõusule (c_1) üldisemat interpretatsiooni. Kui vaatleme ühe sentimeetri võrra pikemaid tudengeid, siis kasvab tudengite keskmine kaal sirge tõusu ehk $0{,}82kg$ võrra. Üldjuhul võiksime öelda järgmist: kui võrdleksime kahte gruppi uuritavaid, kus ühes grupis on x-tunnuse väärtus ühe ühiku võrra suurem kui teises uuritavate grupis, siis oleksid y-tunnuse keskväärtused gruppide vahel c_1 võrra erinevad.

Sageli arvatakse ekslikult, et kui me suurendaksime x-tunnuse väärtust 1 ühiku võrra, siis suureneb y-tunnuse keskväärtus c_1 ühikut. See võib mõningatel juhtudel ka nii olla, aga vaatlusandmete puhul ei pea see enamasti paika: kui võtaksime puntra tudengeid ja venitaksime neid piinapingil 1 cm võrra pikemaks, siis antud piinaprotsedur ei tee neid õnnetuid tudengeid veel 0,82kg võrra raskemateks!

Ka vabaliiget c_0 on vahel võimalik interpreteerida. Kui vaatleme vaid neid uurimisobjekte uuritavas populatsioonis, kelle puhul x-tunnuse väärtus on 0, siis uuritava populatsiooni taoliselt valitud alamhulga y-tunnuse keskväärtust näitabki vabaliige. Kuna aga 0cm pikkust naistudengit ei eksisteeri, siis pole ka antud juhul vabaliikmel sisulist tähendust. Samuti ei tohiks antud sirget kasutada näiteks vastsündinud 50cm pikkuse lapse kaalu leidmiseks: antud regressioonsirge leidmiseks kasutatud tudengite valim võib olla küll esindav kõigi 2. kursuse tudengite jaoks, kuid pole kindlasti esindavaks valimiks vastsündinute jaoks. Näiteks tudengineiude keskmine kaal on 59kg, vastsündinute keskmine kaal on aga kindlasti midagi muud. Arvamine, et tudengite jaoks leitud regressioonsirge võiks sobida vastsündinute kirjeldamiseks, on sama absurdne, kui arvata, et tudengite keskmine kaal kirjeldab vastsündinute keskmist kaalu. Regressioonmudelite kasutamisel tuleb hoolega jälgida, et uuritav, kelle y-tunnuse väärtust soovime prognoosida, kuuluks ikka samasse uuritavasse populatsiooni mida kirjeldab kasutatav regressioonmudel.

On väheusutav, et ühe konkreetse tudengi kaal oleks täpselt võrdne tudengite keskmise kaaluga. Kui tahame kirja panna mudelit juhuslikult valitud tudengi kaalu jaoks, siis peame arvestama, et tema kaal erineb rohkem või vähem keskväärtusest. Isiku eripära, tema kaalu erinevust keskväärtusest (hälvet) tähistatakse sageli sümboliga ε:

$$kaal = -78{,}46 + 0{,}82 \cdot pikkus + \varepsilon. \tag{8.3}$$

Kui varem kirja pandud regressioonmudel (8.2) kirjeldas, kuidas kaalu kesk-

väärtus muutub erineva pikkusega tudengitel, siis regressioonmudel (8.3) kirjeldab, kuidas (juhuslikult valitud) tudengineiu kaal sõltub pikkusest.

Valimis on mõõdetud paljude uuritavate y-tunnuse väärtused. Kui soovime rääkida näiteks i. tudengi kaalust, siis võime kasutada ka alaindekseid:

$$kaal_i = -78{,}46 + 0{,}82 \cdot pikkus_i + \varepsilon_i.$$

Nii on i. tudengil oma isiklik kaal ($kaal_i$), pikkus ($pikkus_i$) ja hälve ehk erinevus keskmisest (ε_i).

Keskmine erinevus keskmisest peab muidugi olema null ($\mathrm{E}\varepsilon = 0$), sest muidu poleks keskmine enam keskmine. Vahel aga tekib vajadus täpsustada, kui suured ja millise jaotusega võivad olla hälbed ehk üksikindiviidide erinevused keskmisest. Sageli eeldatakse (ja tihti ka on) hälbed normaaljaotusega, $\varepsilon \sim N(0; {\sigma_\varepsilon}^2)$. Sellisel juhul on ka y-tunnuse jaotuseks normaaljaotus. Tudengite kaalu-pikkuse näite korral saame näiteks hälvete normaaljaotust eeldades kirja panna, millise jaotusega on mingi kindla pikkusega tudengite kaalud:

$$kaal \sim N(-78{,}46 + 0{,}82 \cdot pikkus; {\sigma_\varepsilon}^2),$$

või üldisema juhu tarvis kirjapandult:

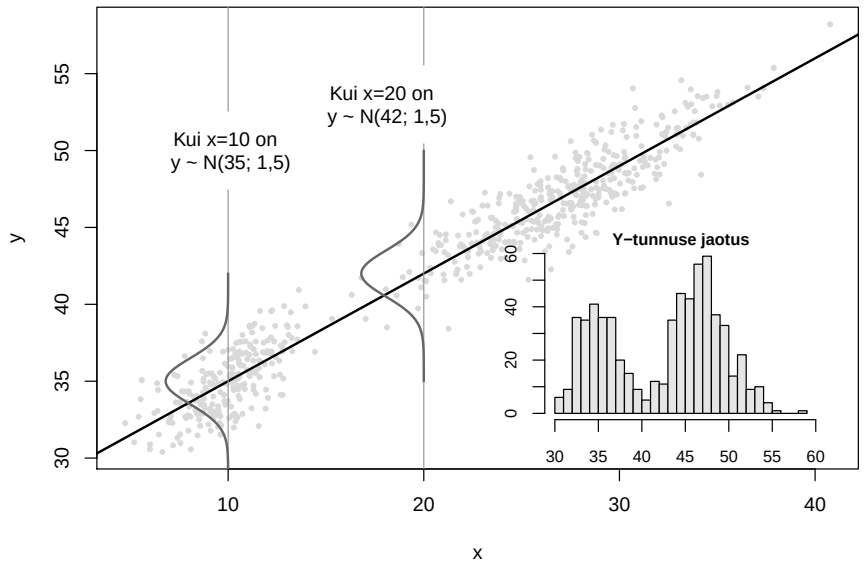
$$y \sim N(c_0 + c_1 x; {\sigma_\varepsilon}^2) \tag{8.4}$$

Taolises olukorras kiputakse rääkima, et uuritav tunnus on normaaljaotusega (näiteks kaal on normaaljaotusega). Paraku tekib siin oht segi ajada kahte täiesti erinevat asja. Regressioonanalüüsi juures peetakse silmas, et y-tunnuse väärtused mingi fikseeritud x-tunnuse väärtuse korral on normaaljaotusega. Näiteks 160cm pikkuste tudengite kaalud on normaaljaotusega. Samuti eeldatakse, et 170 cm pikkuste tudengite jaotus on normaaljaotusega. Aga need kaks normaaljaotust on erinevad, erinevate keskväärtustega. Vahel võib esineda olukordi, kus y-tunnuse jaotus iga konkreetse x-tunnuse väärtuse korral on normaaljaotusega, kuid y-tunnuse enda jaotus (üle erinevate x-tunnuse väärtuste) siiski pole normaaljaotusega. Ka sellisel juhul räägime, et regressioonanalüüsi jäägid on normaaljaotusega ja normaaljaotuse eeldus on täidetud. Taolist võimalikku olukorda kirjaldab joonis 8.3.

Tuleb silmas pidada, et regressioonmudel (8.4) ei kirjelda ära mitte ainult seda, kuidas y-tunnuse keskväärtus x-tunnuse väärtuste muutudes muutub, vaid kirjeldab ära ka selle, kui kaugele keskväärtusest üksikvaatlused sattuda võivad.

Nagu iga mudel, on ka kaalu ja pikkuse omavahelist seost kirjeldav regressioonanalüüsi mudel (veidi) vale. See ei tähenda, et antud mudelist ei võiks

Joonis 8.3: Normaaljaotusega jääkidega regressioonanalüüsi mudel 8.4



olla kasu pikkuse ja kaalu omavahelise seose mõistmiseks või uute, valimis mitteesindatud tudengite kaalu prognoosimisel pikkuse abil. Samas tuleks siiski kontrollida, kas lineaarne seos vähemalt ligikaudseltki sobib kirjeldama seost y ja x tunnuste vahel. Mudeli sobivuse kontrollimise võimalusi on lähemalt kirjeldatud alapeatükis **??**.

8.3 Hinnang ja tema täpsus

Uuritavas populatsioonis võib ju y-tunnuse keskväärtus sõltuda x-tunnuse väärtustest lineaarselt, $\mathrm{E}Y = c_0 + c_1 x$, kuid tavaliselt, vähemalt eluteadustes, pole võimalik kordajate c_0 ja c_1 väärtust võimalik leida teooriale toetudes. Nimetatud tundmatud kordajate c_0 ja c_1 hinnangud $\hat{c}_0$ ja $\hat{c}_1$ tuleb leida valimi põhjal. Valimi põhjal tehtud hinnang on aga paratamatult ekslik, seega tuleb kirjeldada ka hinnangu täpsust.

Esmalt hinnangust endast. Tavaliselt hinnatakse parameetrid c_0 ja c_1 vähimruutude meetodil, st minimiseeritakse olemasolevate vaatluste "prog-

noosimisel" tekkivate prognoosivigade ruutude summa:

$$\hat{c}_0, \hat{c}_1 = \underset{c_0,c_1}{\operatorname{argmin}} \left(y_i - (c_0 + c_1 x)\right)^2 .$$

Miks eelistatakse vähimruutude meetodit? Saksa teadlane Gauss tõestas nimelt paar sajandit tagasi teoreemi, mida tänapäeval tuntakse Gauss-Markovi teoreemi nime all. Antud teoreem väidab järgmist:
Kui

- tegemist on juhusliku valimiga;
- tegelikult kehtib seos $\mathrm{E}Y = c_0 + c_1 x$;
- kui jäägid on sama dispersiooniga iga x väärtuse korral, $\mathrm{D}\varepsilon = {\sigma_\varepsilon}^2 \ \forall x$

siis
on vähimruutude meetodil saadud hinnangud $\hat{c}_0$ ja $\hat{c}_1$ kõige täpsemad nihketa hinnangud parameetritele c_0 ja c_1. Lisaks on parameetrite c_0 ja c_1 mistahes lineaarkombinatsioonile (näiteks suurusele $c_0 + c_1 x$) kõige täpsemaks nihketa hinnanguks vähimruutude meetodil saadud hinnanguid kasutav lineaarkombinatsioon $(\hat{c}_0 + \hat{c}_1 x)$. Märkus: inglise keeles tuntakse sellist hinnangut nime all *BLUE - Best Linear Unbiased Estimator*.

Järgnevalt mõned selgitused teoreemi sõnastuse kohta.

Nihketa hinnang tähendab keskmiselt õiget hinnangut, st ühe valimi korral võime saada kordaja c_1 hinnangu liiga suure ja mõne teise valimi korral liiga väikese, aga keskmiselt, üle kõikmõeldavate valimite, annab hinnangute keskmine kokku parameetri õige väärtuse.

Mida tähendab täpsem antud teoreemi seisukohast? See ei tähenda, et mõni teine meetod ei võiks mõne konkreetse valimi korral anda täpsemat tulemust. See tähendab seda, et kui võtaksime palju juhuslikke valimeid, ning iga valimi põhjal hindaksime meid huvitavad parameetrid, siis ükski teine meetod ei anna väiksema dispersiooniga (väiksema varieeruvusega) hinnanguid. Kuna nihketuse nõudest tulenevalt peavad kõik hinnangud kõikuma õige parameetri väärtuse ümber, siis väike hinnangute varieeruvus tähendab ühtlasi seda, et nad peavad (enamasti) olema lähedal hinnatava suuruse tõelisele väärtusele.

Tavaliselt leitakse regressioonmudeli parameetrite hinnangud arvuti abil. Samas pole nende leidmine ka eriti keeruline, sestap toome siin igaks juhuks ära ka nn käsitsi arvutamiseks kõlvulikud valemid ja vihje nende tuletuskäigule.

Mudeli jääkide (ehk prognoosivigade) ruutude summa (mida minimiseeritakse) on

$$f(c_0, c_1) = \sum_{i=1}^{n} (y_i - c_0 - c_1 x_i)^2.$$

Seda summat käsitletakse tundmatute parameetrite c_0 ja c_1 funktsioonina. Selleks, et leida, millised c_0 ja c_1 väärtused minimiserivad jääkide ruutude summa, tuleb leida funktsiooni $f(c_0, c_1)$ tuletised c_0 ja c_1 järgi ja võrdsustada need nulliga. Saadud võrrandisüsteemi

$$\begin{cases} \frac{\partial f(c_0,c_1)}{\partial c_0} = 0 \\ \\ \frac{\partial f(c_0,c_1)}{\partial c_1} = 0 \end{cases}$$

lahendiks (vajaik algebra pole keeruline, võid ise ka proovida tuletuskäiku läbi teha!) on

$$\begin{aligned} \hat{c}_1 &= \frac{n \sum_{i=1}^n y_i x_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n {x_i}^2 - \left(\sum_{i=1}^n x_i\right)^2} \\ \hat{c}_0 &= \frac{\sum_{i=1}^n y_i - \hat{c}_1 \sum_{i=1}^n x_i}{n} \end{aligned}$$

Muidugi, tõeline matemaatik kontrolliks täiendavalt, kas vaadeldaval funktsioonil on kohas $\hat{c}_0, \hat{c}_1$ ikkagi miinimum (seda ta on) ja mitte näiteks maksimum (kui võrdsustame funktsiooni tuletise nulliga, võime tulemuseks saada kas funktsiooni maksimumi, miinimumi või käänukoha...).

Milleks selline piinarikas kõrvalepõige matemaatika valda? Selleks, et mõista: matemaatikud/statistikud muretsevad tundmatute parameetrite c_0 ja c_1 väärtuste leidmise pärast. Regressioonmudel (8.1) on otsitavate parameetrite c_0 ja c_1 lineaarne funktsioon. Sellepärast kutsutakse antud regressioonmudelit ka lineaarseks (regressioon)mudeliks. See, kas seos tunnuste X ja Y vahel on lineaarne, ei oma tegelikult mingit tähendust. Näiteks mudeli

$$\mathrm{E}Y = c_0 + c_1 x^3$$

puhul on tegemist lineaarse (regressioonanalüüsi) mudeliga, mudeli

$$\mathrm{E}Y = c_0 + c_1^2 x$$

puhul aga pole tegemist lineaarse regressioonimudeliga (sest tegemist on parameetri c_1 ruutfunktsiooniga).

Leitud hinnangud $\hat{c}_0, \hat{c}_1$ on paraku justnimelt hinnangud, ehk veidi rohkem või vähem valed. Kui erinevaid hinnanguid me erinevates (sama suurusega) juhuslikes valimites võiksime näha, seda kirjeldab (hinnangu) standardviga ehk hinnangu standardhälve.

Kuna lineaarse regressioonmudeli parameetrite hinnangute ja nende standardvigade arvutus on juba üsnagi arvutusmahukas tegevus, siis kasutatakse nende leidmiseks enamasti arvutite abi. Alljärgnevalt toetume siingi ühe statistikapaketi (R) väljundile edaspidistes arutlustes. Ka teiste statistikapakettide abil on võimalik teha sarnaseid arvutusi ja jõuda samasuguste tulemusteni.

Alljärgnevalt vaatame väljavõtet statistikapaketi R poolt lineaarse regressioonimudeli kohta trükitavast väljundist:

```
> mudel=lm(kaal~pikkus); summary(mudel)
[...]
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -78.45948    9.42448  -8.325 7.92e-16 ***
pikkus        0.81972    0.05609  14.615  < 2e-16 ***

Residual standard error: 7.298 on 506 degrees of freedom
[...]
```

Programmi R abil leitud hinnang — kuidas tudengi kaal sõltub pikkusest — on samasugune nagu juba varem sai kirja pandud: $kaal = -78{,}46 + 0{,}82 \cdot pikkus + \varepsilon$. Näeme, et tunnuse *pikkus* ees olev kordaja hinnangu $\hat{c}_1 = 0{,}8197$ standardvea hinnang on 0,056, ehk $s(\hat{c}_1) = 0{,}056$. Kui palju erinevaid teadlaseid võtaksid täpselt samasuured juhuslikud valimid Tartu Ülikooli (nais)tudengitest, ning kui kõik need teadlased hindaksid lineaarse regressioonmudeli abil, kuidas sõltub tudengite keskmine kaal tudengi pikkusest, siis nende teadlaste poolt nähtud regressioonsirgete tõusude (hinnangute $\hat{c}_1$) standardhälve oleks ligikaudu 0,056. Saab näidata, et (erinevate teadlaste poolt leitud) hinnangute $\hat{c}_1$ jaotuseks on ligikaudselt normaaljaotus (kui tegemist on kas suure valimiga või kui regressioonmudeli jäägid on normaaljaotusega). Normaaljaotuse puhul jääb aga 95% juhusliku suuruse väärtustest keskväärtusest vähem kui kahe standardhälbe kaugusele. Kuna vähimruutude meetod tagab regressiooniparameetritele nihketa hinnangu, siis saame rääkida, et 95% teadlastel on nende poolt leitud regressioonisirge tõusu hinnang $\hat{c}_1$ vähem kui kahe standardvea kaugusel tõelisest väärtusest c_1. Ka meie poolt vaadeldud hinnangu (0,82) kohta ei oska me öelda, kas

ta üle- või alahindab tegelikku regressioonsirge tõusu (seda, mida näeksime, kui küsitleksime lõpmatult palju tudengeid), kuid võime olla üsna kindlad, et erinevus tegeliku ja meie poolt nähtud numbri vahel ei tohiks olla suurem kahest standardveast. Samasuguse arutluskäiku võime soovi korral korrata ka vabaliikme c_0 kohta.

Kui soovime matemaatiliselt korrektselt leida 95% usaldusintervalli parameetri c_1 (või c_0) tegelikule väärtusele, peame arvestama, et kasutatav standardviga on samuti hinnang (ning võib olla seetõttu ka ise veidi vigane). Korrektse usaldusintervalli leidmiseks peame seetõttu kasutama t-jaotuse kvantiile normaaljaotuse kvantiilide asemel (mistõttu venib usaldusintervall veidi laiemaks). Matemaatiliselt korrektne arvutusvalem leidmaks $(1-\alpha)$-usaldusintervalli parameetrile c_1 on

$$\hat{c}_1 + t_{\alpha/2;df}s(\hat{c}_1) \ldots \hat{c}_1 + t_{1-\alpha/2;df}s(\hat{c}_1),$$

kus df, t-jaotuse vabadusastmete arv (regressioonmudeli kontekstis tuntud ka kui jääkide vabadusastmete arv), leitakse valemiga:

$$\begin{aligned} df &= \text{vaatluste arv} - \text{regressioonmudeli parameetrite arv} \\ &= n - 2 \end{aligned}$$

Esitatud valemit tuleb veidi kommenteerida ja täpsustada: lihtsa regressioonmudeli korral on tegemist kahe hinnatava parameetriga, c_0 ja c_1, seega lahutatakse vaatluste arvust maha 2. Keerukamate mudelite korral, kus uuritava tunnuse (kesk)väärtuse muutumist kirjeldatakse rohkemate tunnuste ja enamate parameetrite abil kasvab ka mahalahutatav number. Kuigi enamasti hinnatakse regressioonmudeli hindamisel ka jääkide hajuvus σ_ε, siis seda (hajuvus)parameetrit vabadusastmete arvutamisel arvesse ei võeta — arvesse lähevad vaid keskväärtuse muutumist kirjeldavad ja vaatluste põhjal hinnatud parameetrid.

Enamasti trükivad statistikapaketid ka regressioonmudeli väljundisse vajaliku vabadusastmete arvu ära (*error degrees of freedom*). Viimast numbrit on mõistlik vaadata kontrollimaks, ega arvuti ja kasutaja vahel pole tekkinud mittemõistmist — kas vabadusastmete arv ikka on kooskõlas vaatluste arvuga, kas ülaltoodud vastavus ikka kehtib?

Näide: Leiame täpse 95% usaldusintervalli pikkuse ees olevale kordajale. Antud regressioonmudeli hinadamisel on kasutatud 508 naistudengi pikkuseid ja kaale. Seega $df = 506$ ja tabelist (või arvutist) tuleb leida t-jaotuse kvantiilid $t_{0{,}025;df=506} = -1{,}96466...$ ja $t_{0{,}975;df=506} = 1{,}96466....$ Edasi läheb

juba lihtsalt:

$$\begin{aligned} 0{,}81972 - 1{,}96466 * 0{,}05609 \quad &\ldots \quad 0.81972 + 1{,}96466 * 0{,}05609 \\ 0{,}7095 \quad &\ldots \quad 0{,}9299 \end{aligned}$$

Ehk 95% kindlusega võime väita, et tegelik regressioonsirge tõus (mille saaksime, kui oleksime mõõtnud üle kõigi Tartu Ülikoolis õppivate tudengineiude kaalud-pikkused) on vahemikus $0{,}7095 \ldots 0{,}9299$.

Paljud statistikapaketid oskavad muidugi ka vastavaid arvutusi automaatselt teha, vaata näiteks statistikapaketi R'i poolt saadavat tulemust. Samas võib käsitsi arvutamise oskus osutuda vajalikuks, kui soovitakse usaldusintervalli leida vaid kirjanduses/artiklis avaldatud informatsiooni põhjal.

Leiame hinnatud parameetritele 95%-usaldusintervalli tarkvarapaketi R abil:

```
> confint(mudel)
                 2.5 %      97.5 %
(Intercept) -96.9754049 -59.9435493
pikkus        0.7095254   0.9299087
```

Näeme juba usaldusintervalle vaadates, et pikkuse ees olev kordaja võib tegelikkuses olla valimi põhjal arvutatud hinnangust $\hat{c}_1 = 0{,}82$ mõnevõrra suurem (kuni 0,9299) või väiksem (kuni 0,7095), kuid kindlasti on tegemist positiivse kordajaga — pikemate tudengite keskmine kaal on suurem kui lühemate tudengite keskmine kaal.

Sarnane küsimus — kas regressioonsirge tõus c_1 võib tegelikkuses olla ka 0 — tekib üllatavalt sageli. Sestap testitakse ka sageli hüpoteesipaare $H_0 : c_1 = 0$ vs $H_1 : c_1 \neq 0$. Vastavat hüpoteesi saab kergesti kontrollida t-testi abil:

$$t := \frac{\hat{c}_1}{s(\hat{c}_1)} \overset{H_0}{\sim} t_{df}$$

ehk parameetri hinnang jagatud tema standardveaga peab nullhüpoteesi kehtides olema t-jaotusega juhuslik suurus. T-jaotuse vabadusastmete arv on jälle see nn jääkide vabadusastmete arv, mida saab leida valemiga (8.5). Edasi saab jätkata nii nagu hüpoteeside statistilisel testimisel tavaks — vaatame, kas saadud t-statistiku väärtus on selline, mida võiksime näha nullhüpoteesi kehtides (kas jääb t-jaotuse 0,025- ja 0,975-kvantiili vahele), või arvutame p-väärtuse ehk tõenäosuse näha sedavõrd ekstreemset (või veel ekstreemsemat) t-statistiku väärtust nullhüpoteesi kehtides. Enamik statistikapakette teeb ülaltoodud hüpoteesidepaari kontrolli ära automaatselt ja

väljastab kohe ka p-väärtuse. Varemtoodud R-programmi väljundist võime näha, et hüpoteesipaari $H_0 : c_1 = 0$ vs $H_1 : c_1 \neq 0$ kontrollimisel saadi p-väärtuseks suurus, mis oli väiksem kui 0,0000000000000002 (<2e-16) ja hüpoteesipaari $H_0 : c_0 = 0$ vs $H_1 : c_0 \neq 0$ kontrollimisel saadi p-väärtuseks 7,92e-16 (0,000000000000000792). Igal juhul on p-väärtused väiksemad kui 0,05 (tüüpiliselt kasutatav olulisustõenäosus) ja mõlemad nullhüpoteesid tuleb kummutada (pole võimalik, et $c_1 = 0$, samuti pole võimalik, et $c_0 = 0$.

Kui regressioonsirge tõus võib uuritavas populatsioonis olla ka 0, siis on võimalik, et Y-tunnuse väärtuste prognoosimiseks kasutatav tunnus X on tegelikult täiesti kasutu, ei sisalda tegelikult informatsiooni Y-tunnuse kohta.

Vahel võib esile kerkida soov kontrollida veidi keerukamaid küsimusi. Näiteks võib mõnikord tekkida vajadus kontrollida, kas kehtib hüpotees $H_0 : c_1 = 1$, või veel üldisemal kujul kirjapandult, võime soovida kontrollida hüpoteesipaari $H_0 : c_1 = c_{H_0}$ (vs $H_1 : c_1 \neq c_{H_0}$), kus c_{H_0} on mõni meie uurimistöö jaoks oluline number (näiteks 1). Ka selliseid hüpoteese saab kontrollida analoogse t-testi abil, kasutades teststatistikut

$$t := \frac{\hat{c}_1 - c_{H_0}}{s(\hat{c}_1)},$$

mis taas nullhüpoteesi kehtides peab olema t-jaotusega, $t \overset{H_0}{\sim} t_{df}$. Selliseid teste arvutitarkvara juba automaatselt ei tee ja vajalikud arvutused tuleb teha ise, kasutades arvuti poolt leitud regressioonparameetrite hinnanguid ja nende standardvigu.

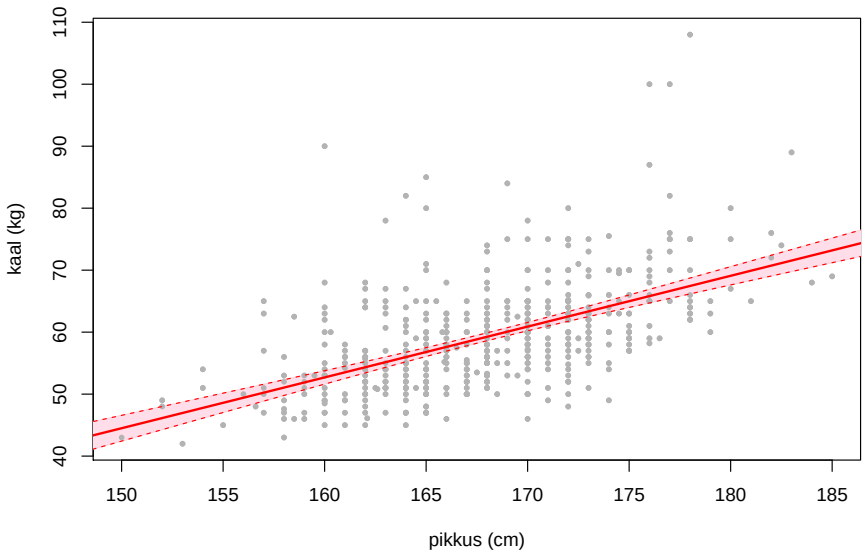
Muidugi tuntakse huvi regressioonmudeli parameetrite ja nende hinnangute täpsuse vastu. Tuleb siiski meeles pidada, et regressioonmudeli parameetrid omavad tähendust vaid seepärast, et ütlevad midagi y-tunnuse kohta, aitavad paremini ja täpsemalt kirjeldada y-tunnuse võimalikke väärtuseid. Sestap on ka tähtis mõista, kui täpselt oskame hinnata y-tunnuse keskväärtust ühel või teisel juhul. Kui täpselt teame, milline on y tunnuse keskväärtus siis, kui vaatleme vaid uuritavaid, kelle x tunnuse väärtuseks on näiteks 165? Vajalikku arvutust pole kahjuks võimalik teha kasutades vaid parameetrite hinnanguid ja nende standardvigu (vaja läheb parameetrite hinnangute kovariatsioonimaatriksit). Õnneks paljud statistikapaketid võimaldavad vajalikke arvutusi teha, nii ka R. Näiteks on võimalik arvutada, kui täpselt me teame 165cm pikkuste tudengineiude kaalu keskväärtust. 95%-usaldusintervalli 165cm pikkuste tudengineiude kaalu keskväärtusele saab statistikapaketis R küsida järgmiselt (kasutame varem hinnatud regressioonmudelit *mudel*):

```
> predict(mudel, data.frame(pikkus=165), interval="confidence")
       fit      lwr      upr
1 56.79383 56.08023 57.50743
```

Vastuseks saame, et hinnang $165cm$ pikkuste tudengineiude kaalu keskväärtusele on $56{,}79\,(-78{,}46 + 0{,}8197 \cdot 165)$, 95%-usaldusintervall keskväärtusele aga $(56{,}08 \ldots 57{,}51)$. Võime olla üsna kindlad, et isegi kui kaaluksime ära kõik 165cm pikkused tudengineiud, siis saadud kaalude keskmine jääks mainitud vahemikku.

Taolisi usaldusintervalle saab leida erinevate pikkuste, erinevate x-tunnuse väärtuste jaoks. Kui kannaksime leitud usaldusintervallid graafikule, tekiks regressioonsirge ümber tema täpsust kirjeldav "koridor". Inglise keeles tuntakse leitut kui *95% pointwise confidence interval*. Vaata ka joonist 8.4, kus hajuvusgraafikule on lisatud regressioonisirge koos usaldusintervalliga.

Joonis 8.4: Regressioonsirge koos 95%-usaldusintervalliga.



NB! Usaldusintervalli interpretatsioon (näitab, kui täpselt teame Y-tunnuse keskväärtust mingi konkreetse X-tunnuse väärtuse korral) kehtib ainult siis, kui Y-tunnuse keskväärtus tõepoolest muutub nii, nagu kasutatud regressioonmudel ütleb. Usaldusintervall arvestab/kirjeldab vaid parameetrite (ebatäpsest) hindamisest tingitud võimaliku vea suurust, mitte aga mudeli valest

valikust tingitud vea suurust!

Kui valitud mudel on aga vale (enamasti on, vähemalt veidikenegi), siis võib leitud usaldusintervallile siiski anda ligikaudse interpretatsiooni: kui hindaksime üldkogumis, kõikmõeldavate uurimisobjektide andmeid kasutades, oma (vigase) regressioonmudeli, siis millise sirgeni võiksime sellisel juhul jõuda? Lisanduvate andmete tõttu saaksime tulemuseks arvatavasti veidi teistsuguse regressioonisirge, kuid üsna kindlasti jääks ta siiski 95%-usaldusintervalli poolt määratud koridori.

8.4 Prognoos ja tema täpsus

Kui on tarvis prognoosida uue objekti Y-tunnuse väärtust, siis pakutakse enamasti prognoosiks Y-tunnuse keskväärtust (keskväärtuse hinnangut). Lisainformatsiooni olemasolul (teame, et X-tunnuse väärtus oli x) pakume prognoosiks nn tinglikku keskväärtus (tema hinnangut): hinnangut nende objektide Y-tunnuse keskväärtusele, kellel $X = x$. Taolise tingliku keskväärtuse saame lihtsalt arvutada enda poolt hinnatud regressioonmudeli abil. Kasutades tudengite kaalu ja pikkuse vahelist seost kirjeldavat regressioonmudelit võime hinnata näiteks 180cm pikkuste tudengite keskmist kaalu. Saadud keskmine sobiks ka ühe uue (varem mittemõõdetud) 180cm pikkuse tudengi kaalu mõistlikuks prognoosiks:

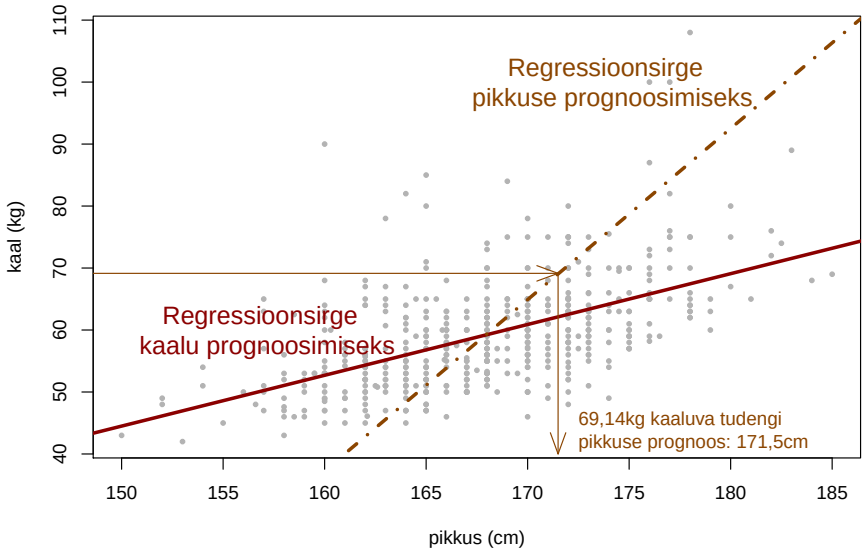
$$\hat{\mathrm{E}}(kaal|pikkus = 180cm) \quad = \quad -78{,}46 + 0{,}82 \cdot 180 = 69{,}14kg$$

Tasub ehk märkida, et 69,14kg raskuse tudengi pikkuse parim prognoos pole 180cm. Kui leiaksime regressioonmudeli, mis kirjaldab, kuidas pikkus sõltub kaalust (antud andmeid kasutades tuleb selleks regressioonmudel $\hat{\mathrm{E}}pikkus = 146{,}5 + 0{,}362 \cdot kaal$), siis saaksime 69,14kg kaaluva tudengi pikkuse prognoosiks hoopis 171,5cm. Kaalu prognoosib pikkuse järgi üks regressioonsirge, aga pikkuse prognoosimiseks kaalu järgi tuleb kasutada täiesti teistsugust regressioonsirget, vaata ka joonist 8.5. Tugevalt soovitatav on luua regressioonmudel just selle tunnuse jaoks, mille väärtuseid on tegelikult vaja prognoosida.

Saadud prognoos võib olla küll parim, mida olemasolevate andmete ja informatsiooni kasutades on võimalik pakkuda, kuid täiesti täpne ta (vähemalt enamasti) pole. Mitte kõik 180cm pikkused tudengid ei kaalu täpselt samapalju. Tahaksime teada, kui suur võib olla leitud prognoosi viga või võimalik eksimus.

Kui täpselt me keskväärtust tegelikult teame, võisime kirjeldada näiteks usaldusintervalli abil. Kuid isegi siis, kui keskväärtus oleks täpselt teada (kõik

Joonis 8.5: Pikkust kaalu abil prognoosiv regressioonsirge ja kaalu pikkuse järgi prognoosiv regressioonsirge



eestimaal elavad inimesed oleksid kaalutud-mõõdetud), poleks võimalik täpselt ennustada järgmise tänaval vastutuleva inimese kaalu. Iga inimene (või uuritav) on alati mõnevõrra individuaalne, tal on oma eripära, mistõttu tema uuritava tunnuse väärtus võib olla mõnevõrra erinev uuritava populatsiooni keskmisest.

Kui erinevad võivad uuritavad olla (tinglikust) keskmisest, seda saab mõõta/kirjeldada standardhälbe abil. Regressioonanalüüsi puhul huvitutakse enamasti regressioonmudeli jääkide standardhälbest (*residual standard error*). Jääkide standardhälve iseloomustab, kui kaugele (tinglikust) keskväärtusest võivad vaatlused tegelikult sattuda (enamasti jäävad vaatlused vähem kui kahe standardhälbe kaugusele keskväärtusest).

Juhul, kui regressioonmudeli jäägid — kõrvalekalded (tinglikust) keskväärtusest — on normaaljaotusega, on võimalik väga täpselt iseloomustada vahemikku, kuhu peaks jääma uue, prognoositava, objekti y-tunnuse väärtus. Seda saab teha prognoosiintervalli abil (peatükk 5.1). Prognoosiintervall iseloomustab vahemikku, kuhu uus, juhuslikult uuritavast populatsioonist valitud vaatlus, jääb soovitud tõenäosusega. Näiteks 95%-prognoosiintervall on vahemik, kuhu prognoositava suuruse tegelik väärtus jääb tõenäosuse-

ga 0,95. Kui regressioonmudeli jäägid on normaaljaotusega (ja valitud regressioonmudel sobib y-tunnuse keskväärtuse muutumise kirjeldamiseks), siis jääb uue, tulevase, vaatluse y tunnuse väärtus 95% tõenäosusega kahe jääkide standardhälbe σ_ε kaugusele regressioonsirgest ehk y-tunnuse tinglikust keskväärtusest. Prognoosiintervalli täpseks arvutamiseks tuleb arvesse võtta ka seda, et tegelikku jääkide standardhälvet pole teada (kasutada saab ainult jääkide standardhälbe hinnangut), samuti ka seda, et ka regressioonsirge on kõigest hinnang ning regressioonsirge parameetrid võivad olla hinnatud veaga. Regressioonmudelit kasutades leitud prognoosile prognoosiintervalli leidmine on enamasti arvutuslikult tülikas, sestap on mugavam lasta vajalikud arvutused teha arvutil. Enamikus statistikapakettides on vastav võimalus olemas, R-is saab prognoosiintervalli leida predict-käsu abil. Näiteks 95%-prognoosiintervalli uue tudengi kaalule, kelle pikkus on 165cm, saab leida nii:

```
> predict(mudel, data.frame(pikkus=165), interval="prediction")
       fit      lwr      upr
1 56.79383 42.43865 71.14901
```

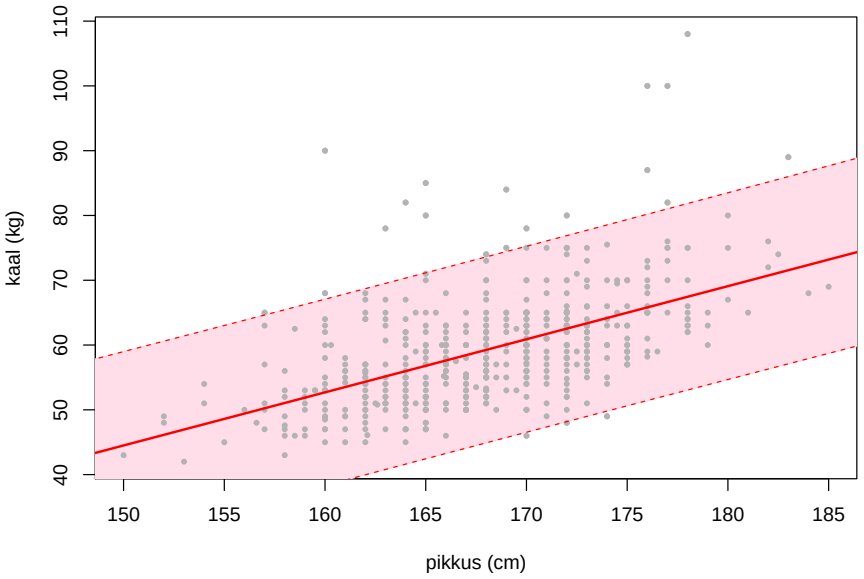
Selgub, et 165cm pikkuse tudengi kaaluks prognoosib kasutatav regressioonmudel 56,8kg, kuid 95% tõenäosusega jääb uue (valimisse mittekuulunud tudeng, selline tudeng, kelle pikkust ja kaalu regressioonmudeli hindamisel ei kasutatud) 165cm pikkuse tudengineiu kaal vahemikku 42,4kg kuni 71,2kg. Märkus: usaldusintervallide ja prognoosiintervallide puhul on heaks tavaks, et intervalli ümmardatakse laiemaks, näiteks: 42kg..72kg.

Taolisi prognoosiintervalle saab leida muidugi iga mõeldava pikkuse (x-tunnuse väärtuse) jaoks. Saadud prognoosiintervallid saab siis soovi korral ka hajuvusgraafikule kanda, saamaks koridori, kuhu järgmise tudengi kaal peaks suure tõenäosusega (95%) sattuma. Saadud graafikut iseloomustab joonis 8.6. Võrdle ka prognoosiintervalli iseloomustavat joonist usaldusintervalli kirjeldava joonisega (joonis 8.4). Võiks märgata, et usaldusintervall on märkimisväärselt kitsam prognoosiintervallist — kaalu keskväärtust on võimalik hinnata märksa täpsemalt, kui prognoosida üksiku juhuslikult valitud tudengi kaalu!

Mida prognoosiintervalli leidmisealgoritm (statistikapakett) arvestab ja mis jääb kasutaja mureks? Prognoosiintervalli arvutamisel võetakse enamasti arvesse, et regressioonsirge parameetrite hinnangud ei pruugi olla korrektsed ja täiesti õiged (arvestatakse, et $\hat{c}_0 \neq c_0$, $\hat{c}_1 \neq c_1$, $\hat{\sigma}_\varepsilon \neq \sigma_\varepsilon$), arvestatakse ka sellega, et y-tunnus on juhuslik, igal uurimisobjekt on eripärane.

Mis jääb prognoosiintervalli kasutaja vastutusele?

Joonis 8.6: 95%-prognoosiintervall.



- Kõigepealt, kasutaja peab vastutama, et tegelikult muutub Y tunnuse keskväärtus X tunnuse väärtuste muutudes tõepoolest lineaarset, ehk kasutaja peab garanteerima, et regressioonmudel 8.1 tõepoolest kehtib. Ei prognoosiintervall ega usaldusintervall ei üritagi kirjeldada võimalikku viga, mis tuleneb regressioonmudeli valest valikust (kui seos x-tunnuse ja y-tunnuse vahel pole lineaarne, vaid midagi muud, siis prognoosiintervalli ega usaldusintervalli arvutused pole enam korrektsed — kuigi kui mudel on ligikaudu õige, siis *võivad* ka prognoosi- ja usaldusintervallid *ligikaudu* paika pidada).

- Prognoosiintervalli arvutusvalem on väga tundlik normaaljaotuse eelduse suhtes — kui regressioonmudeli jääkide jaotuseks pole normaaljaotus, siis on ka prognoosiintervall valesti arvutatud. Tõsi, kui jääkide jaotus on enam-vähem normaaljaotusega, siis võiks ka leitud prognoosiintervall olla enam-vähem korrektne.

- Prognoosiintervalli arvutamisel eeldatakse, et regressioonmudeli jääkide hajuvus on kogu aeg samasuur — olgu x-tunnuse väärtus kas suur või väike, ikka on y-tunnuse väärtuste võimalikud kõrvalekalded ligi-

kaudu samasuured ehk jääkide dispersioon peaks olema samasugune mistahes x-tunnuse väärtuse korral.

Antud eelduste kontrolli võimalusi tutvustame regressioonmudeli eeldustest rääkivas alampeatükis.

Kui regressioonmudeli jäägid pole normaaljaotusega, siis on võimalik siiski leida ligikaudselt korrektset prognoosiintervalli, kasutades näiteks D.J.Olive (Olive, 2006) poolt pakutud nn jaotusvaba prognoosiintervalli arvutusvalemit. Olive mitteparameetrilise usaldusintervalli arvutusvalem pole kahjuks statistikapaketis R valmiskujul realiseeritud, küll aga on võimalik vajalik arvutus ise väikese vaevaga läbi teha. Alljärgnevalt on ära toodud R-keeles kirja pandud programm, mis leiab ligikaudselt korrektse prognoosiintervalli ilma jääkide normaaljaotust eeldamata (prognoosiintervall uue, 165cm pikkuse, tudengi kaalule):

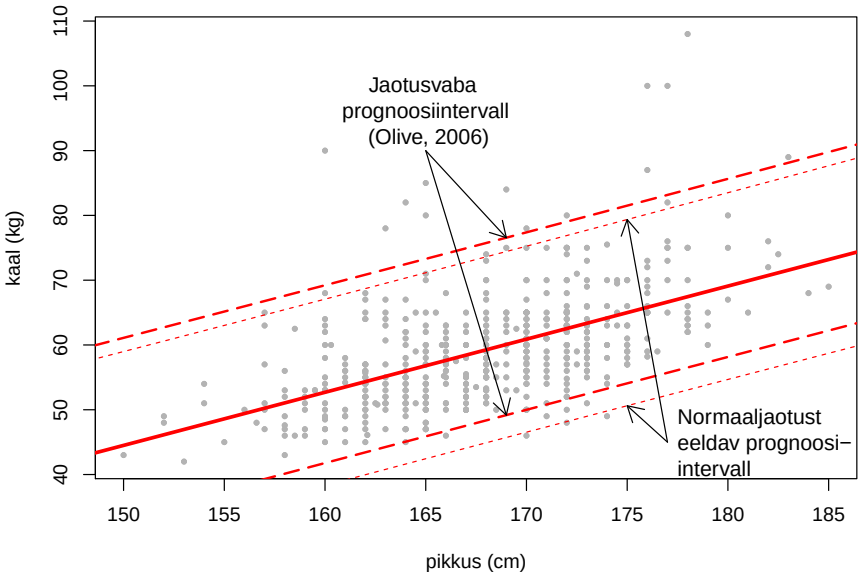
```
# Kasutatav regressioonmudel: prognoosime kaalu pikkuse abil
mudel=lm(kaal~pikkus)

# X-tunnuse väärtus, mille jaoks prognoosiintervalli leiame:
x=165

# Regressioonmudeli hinnatavate parameetrite arv;
# lihtsa regressioonmudeli korral p=2
p=length(coef(mudel))

# Regressioonmudeli hindamiseks kasutatud vaatluste arv
n=mudel$df+p

aa=predict(mudel, data.frame(pikkus=x), interval="confidence")
q_alumine = quantile(residuals(mudel), 0.025)
q_ylemine = quantile(residuals(mudel), 0.975)
se=summary(mudel)$sigma
konst=(1+15/n)*sqrt(n/(n-p)*(1+(aa[,3]-aa[,1])/(se**2*qt(0.975,mudel$df))))

alumine_PI=aa[,1]+konst*q_alumine
ylemine_PI=aa[,1]+konst*q_ylemine

alumine_PI; ylemine_PI
```

Antud programmi tulemusena saaksime 95%-prognoosiintervalliks 45,89...73,29, mis on märgatavalt erinev kui jääkide normaaljaotuse eeldusel leitud prog-

noosiintervall (42,43...71,15). Selline märkimisväärne erinevus erinevatel meetoditel leitud prognoosiintervallide vahel viitab võimalusele, et mõni varemtehtud eeldustest ei pea paika (äkki antud mudeli korral pole regressioonmudeli jäägid ikkagi normaaljaotusega?).

Normaaljaotuse eeldusest hoiduva Olive prognoosiintervalli ja klassikalise, jääkide normaaljaotust eeldava prognoosiintervalli võrdlus on toodud joonisel 8.7.

Joonis 8.7: 95%-prognoosiintervall normaaljaotuse eeldusel ja ilma selleta (nn jaotusvaba prognoosiintervall).

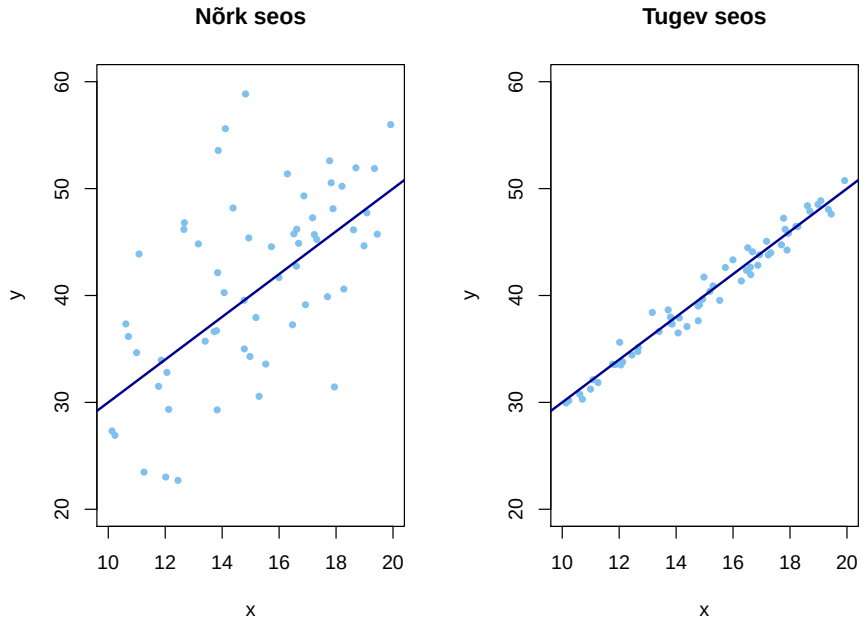


Tuleb siiski meeles pidada, et Olive jaotusvaba prognoosiintervall vabastab meid vaid ühe eelduse painest — jääkide jaotus ei pea enam olema normaaljaotus — kuid kaks järgijäänud eeldust (sama uuritava tunnuse hajuvus kõigi x-tunnuse väärtuste korral ja regressioonmudeli kuju sobivus) jäävad siiski alles.

8.5 Regressioonseose tugevus

Kui eksisteerib statistiline seos tunnuste X ja Y vahel, siis tänu X-tunnuse väärtuse teadmisele oskame midagi täpsemalt öelda ka Y-tunnuse väärtuse kohta. Aga kui palju kasu ikkagi X-tunnuse väärtuse teadmisest on? Kui tugev on seos X-tunnuse ja Y-tunnuse vahel? Kui palju täpsema prognoosi me tänu X-tunnuse väärtuse teadmisele suudame anda Y-tunnuse väärtusele? Vaata näidet tugevast seosest ja nõrgast seosest joonisel 8.8.

Joonis 8.8: Regressioonseose tugevus. Nõrk seos ja tugev seos.



8.5.1 Determinatsioonikordaja R^2

Prognoosi täpsuse kirjeldamiseks on võimalik kasutada regressioonmudeli jääkide ehk prognoosivigade dispersiooni ($\sigma_\varepsilon{}^2$). Võrdluseks võime võtta olukorra, kus pole võimalik kasutada X-tunnuse väärtust prognoosimisel. Sellisel juhul oleks parimaks mõeldavaks uue uurimisobjekti Y-tunnuse prognoosiks lihtsalt Y-tunnuse keskväärtus $\mathrm{E}Y$. Taolise naiivse prognoosi puhul oleks prognoosivigade $\varepsilon_{naiivne} = Y - \mathrm{E}Y$ dispersioon võrdne Y-tunnuse dispersiooniga, $\mathrm{D}(\varepsilon_{naiivne}) = \mathrm{D}(Y - \mathrm{E}Y) = \mathrm{D}Y$ (meenuta dispersiooni 3. oma-

dust). Seega võit prognoosi täpsuses tänu X-tunnuse väärtuse teadmisele on $\mathrm{D}Y - {\sigma_\varepsilon}^2$. Üks võimalus olekski kasutada antud näitajat seose tugevuse iseloomustamiseks, aga... tema tõlgendamine osutub sageli ebamugavaks. Kui muudetaks mõõtühikut, milles mõõdetakse Y-tunnust, muutub ka võit täpsuses (kui kaalu mõõdetakse grammides kilogrammide asemel, muutub saavutatud võit täpsuses 1000000 korda "suuremaks". Seetõttu eelistatakse mõõta suhtelist võitu täpsuses, kui suure osa Y-tunnuse hajuvusest (dispersioonist) oli võimalik ära kirjeldada tänu X-tunnusele. Saadud tulemust esitatakse sageli ka protsendina (mitu protsenti prognoosivigade hajuvus väheneb tänu X-tunnuse väärtuste teadmisele):

$$\frac{\mathrm{D}Y - {\sigma_\varepsilon}^2}{\mathrm{D}Y} \ (\cdot 100\%).$$

Antud valem kirjeldab ideaalnäitajat, mille abil võiks regressioonseose tugevust kirjeldada. Paraku seda ideaalnäitajat pole võimalik niisama lihtsalt arvutada — tema leidmiseks oleks tarvis teada uuritava tunnuse hajuvust populatsioonis ($\mathrm{D}Y$), ja prognoosivigade tegelikku dispersiooni ${\sigma_\varepsilon}^2$. Nende näitajate väärtuseid on aga peaaegu alati teadmata, tuleb kasutada nimetatud näitajate hinnanguid. Kui suuruse $\mathrm{D}Y$ hindamises ollakse enamasti üksmeelel, tema hinnanguks kasutatakse tavalist valimidispersiooni $s^2(Y) = \frac{1}{n-1}\sum (y_i - \overline{y})^2$, siis regressioonijääkide dispersiooni hindamisel esineb kaks üsnagi erinevat võimalust. Ühel juhul hinnatakse regressioonijääkide dispersiooni vägagi klassikalisel viisil — leitakse olemasolevate vaatluste prognoosimisel tehtud prognoosivead ja leitakse siis nende prognoosivigade dispersiooni hinnang kasutades harilikku valimisdispersiooni arvutamise valemit:

$$\begin{aligned} e_i &= y_i - (\hat{c}_0 + \hat{c}_1 x_i) \\ {\tilde{\sigma}_\varepsilon}^2 &= \frac{1}{n-1}\sum (e_i - \overline{e})^2 \end{aligned}$$

Viimast valemit saab veel veidi lihtsustada, sest $\overline{e} = 0$.

Tulemuseks saadakse näitaja, mida kutsutakse determinatsioonikordajaks, R^2:

$$R^2 := \frac{{s_Y}^2 - {\tilde{\sigma}_\varepsilon}^2}{{s_Y}^2} \ (\cdot 100\%)$$

Kui determinatsioonikordajat hakati juba hoolega kasutama, märgati, et antud hinnanguga on seotud üks probleem. Nimelt on suuruse ${\sigma_\varepsilon}^2$ hinnang, mida determinatsioonikordaja arvutamisel kasutatakse (${\tilde{\sigma}_\varepsilon}^2$), nihkega hinnang — ta alahindab (keskmiselt) regressioonimudeli jääkide dispersiooni.

Probleemi alge peitub selles, et regressioonmudeli parameetrid pole täpselt teada, tegemist on kõigest hinnangutega. Kuna regressioonsirge paigutatakse nii, et ta kõige paremini kirjeldaks valimis olemasolevaid vaatluseid, siis saadaksegi sirge, mis väga hästi kirjeldab olemasolevaid, regressioonsirge hindamiseks kasutatud vaatluseid. Uusi, tulevasi vaatluseid ei pruugi aga regressioonsirge enam samavõrra hästi kirjeldada — tulevased vaatlused võivad vigaselt hinnatud regressioonsirgest paikneda kaugemal, kui olemasolevate vaatluste pealt võiks arvata.

Parem, nihketa hinnang jääkide dispersioonile (tulevaste vaatluste keskmine ruutkaugus regressioonisirgest) oleks suurus

$$\hat{\sigma_\varepsilon}^2 \quad = \quad \frac{1}{n-p} \sum (e_i - \overline{e})^2$$

Kus p on regressioonsirge kirjeldamiseks hinnatavate parameetrite arv (lihtsa regressioonmudeli korral $p = 2$, sest hinnatakse vabaliige c_0 ja sirge tõus c_1). Kui kasutatakse mõistlikumat hinnangut jääkide dispersioonile saadakse näitaja, mida kutsustakse kohandatud või parandatud determinatsioonikordajaks (*adjusted* R^2) R^2_{adj}:

$$R^2_{adj} := \frac{{s_Y}^2 - \hat{\sigma_\varepsilon}^2}{{s_Y}^2} \ (\cdot 100\%).$$

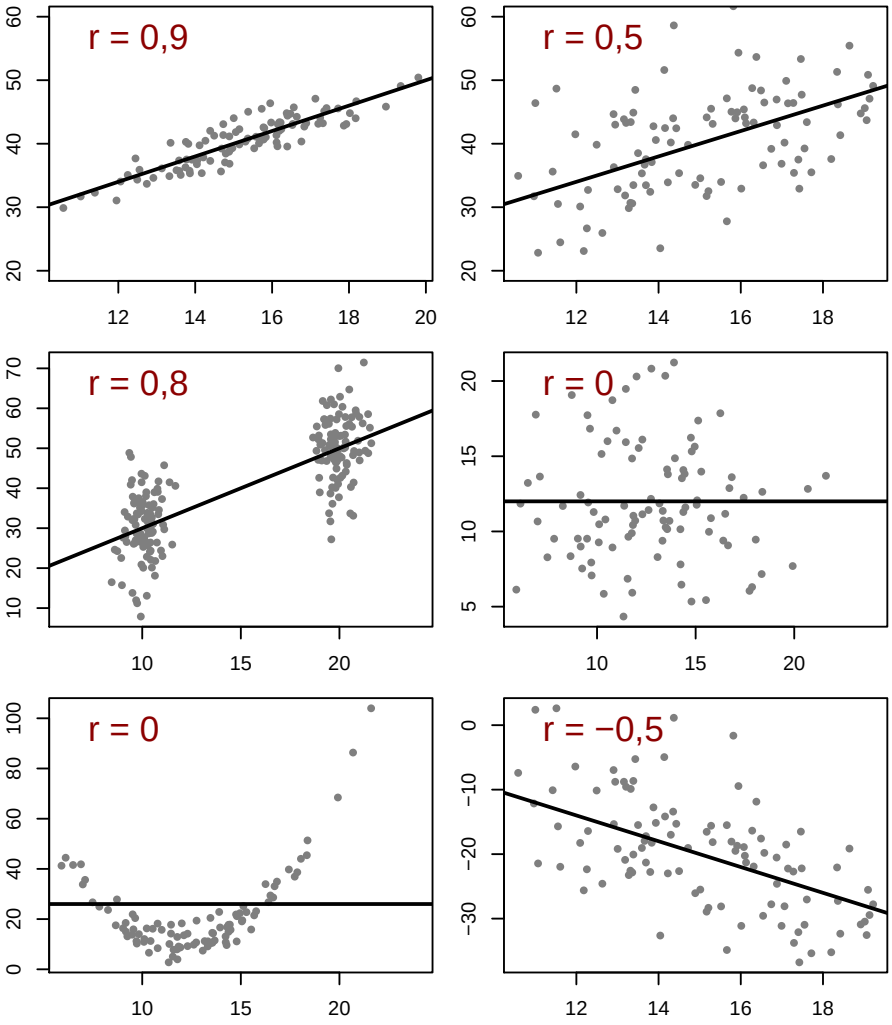
Matemaatiku vaatevinklist on kohandatud determinatsioonikordaja kasutamine paremini õigustatud kui tavalise determinatsioonikordaja kasutamine, ajaloolistel põhjustel kohtab kirjanduses siiski sageli ka tavalist determinatsioonikordajat. Kohandatud determinatsioonikordaja on alati veidi väiksem kui tavaline determinatsioonikordaja. Mõlema näitaja interpretatsioon on sama — mõlemad näitajad iseloomustavad, kui suure osa uuritava tunnuse hajuvusest on võimalik kirjeldada kasutades tunnust X ehk kui palju väheneb Y-tunnuse prognoosivigade dispersioon kui on võimalik kasutada prognoosimisel X-tunnuse väärtuseid (võrreldes olukorraga, kus X-tunnuse väärtuseid pole võimalik Y-tunnuse väärtuste prognoosimisel kasutada).

8.5.2 Lineaarne korrelatsioonikordaja r

Arvatavasti kõige sagedamini kasutatakse kahe tunnuse vahelise seose tugevuse iseloomustamiseks lineaarset korrelatsioonikordajat r (tuntud ka kui Pearsoni korrelatsioonikordaja). Korrelatsioonikordaja ruut on determinatsioonikordaja, kusjuures lineaarne korrelatsioonikordaja on positiivne, kui

ühe tunnuse väärtuste kasvades teise tunnuse väärtused kipuvad samuti kasvama. Kui aga ühe tunnuse väärtuste kasvades teise tunnuse väärtused pigem kahanevad, siis on korrelatsioonikordaja negatiivne. Näiteid erinevatest hajuvusgraafikutest ja nendele vastavatest korrelatsioonikordajatest r vaata jooniselt 8.9.

Joonis 8.9: Pearsoni lineaarne korrelatsioonikordaja. Näiteid.



Pearsoni korrelatsioonikordaja omadused:

- Kui tunnuste X ja Y vahel on lineaarne funktsionaalne seos $Y = c_0 +$

c_1X (ehk täpne lineaarne seos), siis on korrelatsioonikordaja väärtus kas 1 või -1 vastavalt kordaja c_1 märgile.

- Kui $r > 0$, siis ühe tunnuse suurenedes keskmiselt teine tunnus kasvab ja vastupidi — ühe vähenedes väheneb ka teine.
- Kui $r < 0$, siis ühe tunnuse väärtuste suurenedes keskmiselt teise tunnuse väärtused kahanevad ja vastupidi — ühe kahanedes teine kasvab.
- Kui tunnused on lineaarselt sõltumatud (tunnuste vahel võib aga olla mittelineaarne sõltuvus), siis on korrelatsioonikordaja null $r = 0$.
- Korrelatsioonikordaja ruut r^2 ehk determinatsioonikordaja $r^2 = R^2$ näitab, kui suur osa ühe tunnuse hajuvusest (dispersioonist) on kirjeldatud teise poolt (lineaarse regressioonmudeli abil).
- Mida suurem on korrelatsioonikordaja absoluutväärtus, seda tugevam on korrelatiivne seos tunnuste vahel.
- Mõõtühiku (lineaarne) vahetus ei muuda korrelatsioonikordaja suurust (Korrelatsioonikordaja ei muutu, kui mõõdame temperatuuri Celsiuse kraadide C^o asemel Farenheitides F^o, samuti võime pikkust mõõta sentimeetrites või meetrites- korrelatsioonikordaja jääb ikka samaks).

Pearsoni lineaarne korrelatsioonikordaja iseloomustab ainult lineaarse seose tugevust. Kui tunnuste X ja Y vahelist seost ei sobi kirjeldama sirge, siis võib korrelatsioonikordaja r väärtus osutuda ka nulliks või nullilähedaseks isegi siis, kui tegelikult eksisteerib tugev (mittelineaarne) statistiline seos tunnuste vahel.

Determinatsioonikordajat saab edukalt kasutada ka keerukamate mudelite juures (näiteks selliste mudelite juures, kus kasutatakse paljusid erinevaid tunnuseid Y-tunnuse prognoosimiseks) — on ju ikkagi võimalik mõõta, kui palju täpsemaks muutus prognoos tänu kasutada olevale lisainformatsioonile. Seega determinatsioonikordajat nähes võib tegemist olla nii lihtsa regressioonmudeli kui ka keeruka, paljusid erinevaid tunnuseid sisaldava regressioonmudeliga. Kui raporteeritakse Pearsoni lineaarset korrelatsioonikordajat r, siis võib olla üsna kindel, et kasutatud on kõige lihtsamat lineaarset regressioonmudelit.

8.6 Eeldused

Regressioonanalüüsi tehes, regressioonanlüüsi mudelit hinnates, hinnangu täpsust kirjeldades, usaldusintervalle või prognoosiintervalle leides oleme teinud erinevaid eeluseid. Alljärgnevalt vaatleme, milliste eelduste tegemist on tarvis olnud ja kuidas tehtud eelduseid saaks kontrollida.

8.6.1 Juhuslik ja esindav valim

Kui on soov regressioonanalüüsi abil saadud tulemusi üldistada uuritavale populatsioonile (näiteks kasutatakse kas usaldusintervalli, p-väärtust, prognoosiintervalli), siis on see võimalik vaid siis, kui tegemist on juhusliku

Kõigepealt lühike ülevaade erinevatest eeldustest, mis käesolevas peatükis ühes või teises kohas on vajalikuks osutunud.

- **Juhuslik ja esindav valim.**

on erinevate tulemuste saamiseks tehtud erinevaid eelduseid. Kui räägitakse, et üks või teine tunnus on normaaljaotusega, siis eeldatakse, et selle tunnuse histogramm/tihedusfunktsioon on tuttava kellukese kujuga

Levinud viga reSee ei tähenda, et kui võtaksime mõne tudengi ja venitaksime teda piinapingil 1cm võrra pikemaks, et tema kaal sellepärast koheselt 0,82kg suureneks!

Tulemust — prognoosivigade ruutude summat minimiseerivat sirget — võib näha joonisel 8.10.

8.6.2 Kõrvalepõige: miks minimiseeritakse prognoosivigade ruutude summat?

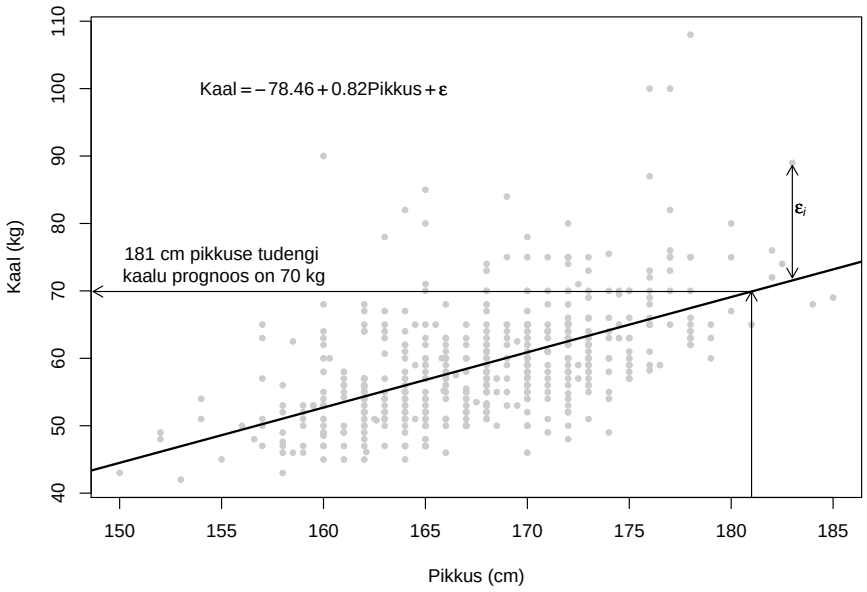
Täpsemini prognoosiva sirge valikuks oleks ka teisi võimalusi. Näiteks võiksime otsida sirget, mille puhul prognoosivigade absoluutväärtuste summa oleks minimaalne. Tulemuseks oleks veidi teistsugune "parim" sirge. Miks siis eelistatakse prognoosivigade ruutude summat?

Selle selgitamiseks vaatame esmalt olukorda, kus vaatlustulemuste $y_1, y_2, \ldots, y_n$ prognoosimiseks pole võimalik kasutada teise tunnuse (x-tunnuse) abi. Milline väärtus $\hat{y}$ osutuks siis uue vaatluse parimaks prognoosiks, kui kasutame vähimruutude meetodit?

Prognoosivigade ruutude summa oleks sellisel juhul

$$(y_1 - \hat{y})^2 + (y_2 - \hat{y})^2 + \ldots + (y_n - \hat{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

Joonis 8.10: Prognoosivigade ruutude summat minimiseeriv sirge



Leidmaks $\hat{y}$ väärtust, mis tagaks väikseima prognoosivigade ruutude summa, leiame esmalt funktsiooni tuletise suuruse $\hat{y}$ järgi (tegutseme nii, nagu ikka funktsiooni miinimumi või maksimumi otsides):

$$\begin{aligned}
\frac{\partial f(\hat{y})}{\partial \hat{y}} &= \frac{\partial}{\partial \hat{y}} \sum_{i=1}^{n} (y_i - \hat{y})^2 \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \hat{y}} (y_i - \hat{y})^2 \\
&= \sum_{i=1}^{n} 2(y_i - \hat{y}) \cdot \frac{\partial}{\partial \hat{y}} (y_i - \hat{y}) \\
&= \sum_{i=1}^{n} 2(y_i - \hat{y})(-1).
\end{aligned}$$

Leitud tuletise võrdsustame nulliga ja lahendame saadud võrrandi:

$$\begin{aligned}
0 &= \sum_{i=1}^{n} 2(y_i - \hat{y})(-1) \\
0 &= \sum_{i=1}^{n} (y_i - \hat{y}) \\
\sum_{i=1}^{n} \hat{y} &= \sum_{i=1}^{n} y_i \\
n\hat{y} &= \sum_{i=1}^{n} y_i \\
\hat{y} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\
\hat{y} &= \overline{y}.
\end{aligned}$$

Koolimatemaatikat mäletav inimene teab nüüd, et valides $\hat{y} = \overline{y}$ me kas maksimiseerime või minimiseerime prognoosivigade ruutude summa (või on tegemist funktsiooni käänukohaga). Võttes teise tuletise $\hat{y}$ järgi prognoosivigade ruutude summast vaatame, kas leitud teine tuletis on positiivne või negatiivne kohal $\hat{y} = \overline{y}$:

$$\begin{aligned}
\frac{\partial^2 f(\hat{y})}{\partial \hat{y}^2} &= \frac{\partial \sum_{i=1}^{n} 2(y_i - \hat{y})(-1)}{\partial \hat{y}} \\
&= \frac{\partial \sum_{i=1}^{n} 2\hat{y}}{\partial \hat{y}} \\
&= 2n.
\end{aligned}$$

Nagu näeme, on vähimruutude meetodi mõttes (minimiseerib prognoosivigade ruutude summat) parimaks prognoosiks valimi keskmine (mis omakorda on hinnang keskväärtusele). Seega, kui minimiseerime prognoosivigade ruutude summat, siis käseb matemaatika pakkuda uue väärtuse prognoosiks keskväärtust (või selle hinnangut, valimikeskmist), vähemalt juhul, kui meil puudub igasugune lisainformatsioon prognoositava objekti kohta. Kui miniseriksime mitte prognoosivigade ruutude summat, vaid näiteks prognoosivigade absoluutväärtuste summat, pakutaks lisainformatsiooni puudumisel meile "parimaks" prognoosiks midagi keskväärtuset/keskmisest erinevat.

Vaatame nüüd veidi keerukamat juhtu — olukorda, kus prognooside tegemiseks saame kasutada ka X-tunnuse abi. Oletame lisaks, et tegelik seos Y tunnuse ja X vahel on tõepoolest lineaarne, st. kirjeldatav kujul

$$Y = c_0 + c_1 X + \varepsilon,$$

kus ε kirjeldab ühe uuritava indiviidi omapära, hälvet keskmisest ($\mathrm{E}\varepsilon = 0$).

Kui vaatame, milline on Y tunnuse keskväärtus neil uurimisobjektidel, kellel $X = x$:

$$\begin{aligned} \mathrm{E}(Y|X = x) &= \mathrm{E}(c_0 + c_1 x + \varepsilon) \\ &= c_0 + c_1 x \end{aligned}$$

Analoogse arvutuse abil saame näidata, et minimiseerides

Sellisel juhul on Y tunnuse keskväärtus $c_0 + c_1 x$ $Y-$

Sellisel juhul oleks vähimruutude meetodil hinnatud sirge $\hat{c}_0 + \hat{c}_1 x$ nihketa hinnanguks tegelikule seosele. Lisaks võiksime siis saadud sirget interpreteerida järgmisel viisil: sellistel uuritavatel, kellel x-tunnuse väärtuseks on x, on y-tunnuse

olukorra juurde. Oletame, et tegelik seos

tuleks siis parim

(uute väärtuste) täpsemate Üsna hoomatav on soov valida sirge selliselt, et olemasolevaid andmeid

puudub aga üldse arusaam tunnustevahelise mida me nende seoste põhjal võime , ligikaudsetest vastustest. Aga vahel meile sedavõrd intuitiivselt haaratavad

lugemist ipad'ile või kindle

Vahel soovime prognoosida pideva tunnuse väärtuseid. Käesolevas peatükis eeldame, et võime prognoosimiseks kasutada üheainsa teise pideva tunnuse väärtuseid. Näiteks võime

Juhul, kui uuritav tunnus on normaaljaotusega, piisab uuritava tunnuse jaotuse määramiseks kui teame, millised on uuritava tunnuse keskväärtus ja dispersioon. Eeldame esialgu, et uuritava tunnuse hajuvus (dispersioon) ei muutu katsetingimuste muutudes. Sellisel juhul piisab uuritava tunnuse jaotuse (muutumise) kirjeldamiseks, kui suudame kirjeldada, kuidas muutub uuritava tunnuse keskväärtus.

teise järgi teist pidevat tunnust.

Huvitagu kedagi näiteks mingite ringide pindalad (näiteks puu ristlõike pindala) ja kuidas tunnus x (näiteks puu vanus) seda pindala mõjutab. Sellisel juhul ei tohiks teha regressioonmudelit kus ringi raadisu r või diameeter d

sõltub tunnusest x. Kui keegi ekslikult arvab, et hästi ringi raadiust prognoosiv mudel võimaldab ka täpselt ringi pindala prognoosida (kasutades näiteks valemit $S = \pi r^2$), siis ta enamasti eksib — otse pindala prognoosimiseks tehtud mudel prognoosib tõepoolest ringi pindalat palju täpsemalt võrreldes keerukama lähenemisega, kus kasutatakse hästi ringi raadiust prognoosivat mudelit ja valemit, mis ütleb, kuidas Kui kedagi näiteks huvitab, kuidas tunnus x mõjutab mingite ringide pindalasid, j

Peatükk 9

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?

9.1 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 9.1.

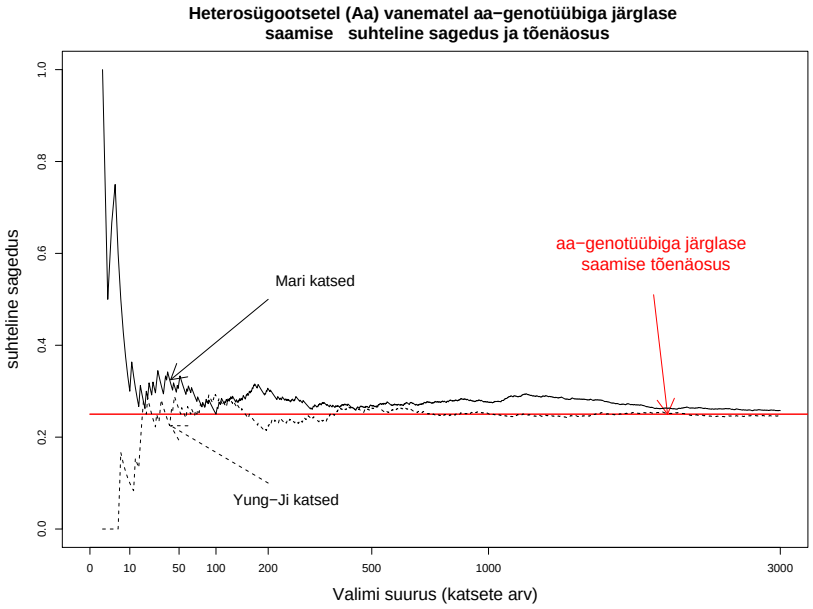
Definitsioon 9.1 *Juhusliku sündmuse toimumise tõenäosuseks $P(A)$ nimetatakse sündmuse A toimumise suhtelist sagedust peale lõpmatult paljude katsete sooritamist. Matemaatiliselt korrektselt kirja pandult:*

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

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 9.1: 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.

9.1.1 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

$$P(A) = k/n, \tag{9.1}$$

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 9.1 *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 (9.1): $P(A) = 2/6 = 1/3$ (sest kõik visketulemused on võrdvõimalikud).*

Näide 9.2 *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 (9.1)? 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 $2/6\,(= 1/3)$ — $P(A) < 1/3$.*

9.2 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?

9.2.1 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:

$$P(X = x) = 0{,}1 \times 0{,}9^{x-1}. \tag{9.2}$$

Sellisel juhul ütleme, et juhusliku suuruse X tõenäosusfunktsioon on antud valemiga (9.2) — 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!).

9.2.2 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).

9.2.3 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 9.3 *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:*

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

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

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

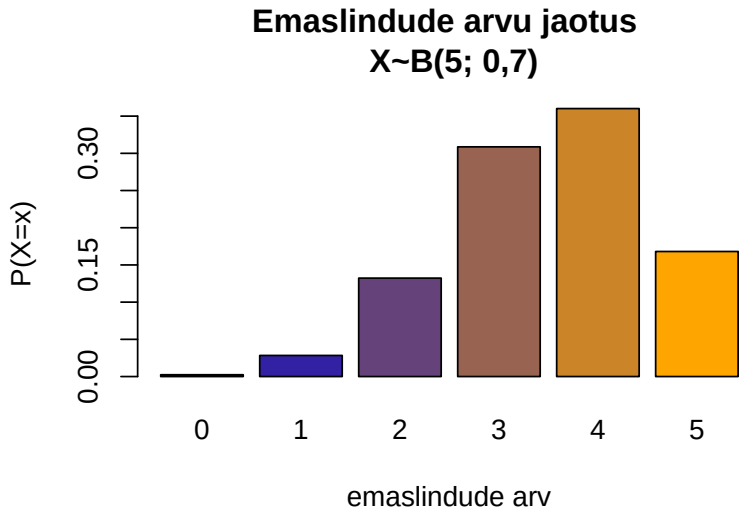
Seega on emaste pleektatsude arvu jaotus, $B(2; 0.6)$, kirja pandav järgmise tabeli abil:

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

Näide 9.4 *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 9.2.



Joonis 9.2: Binoomjaotusega juhusliku suuruse $X \sim B(5; 0.7)$ jaotus

9.2.4 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.
- . . . .

9.2.5 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 9.3, kus on esitatud ühe juhusliku suuruse jaotus- ja tijedusfunktsioon.

Teades tihedusfunktsiooni, on võimalik leida jaotusfunktsiooni:

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

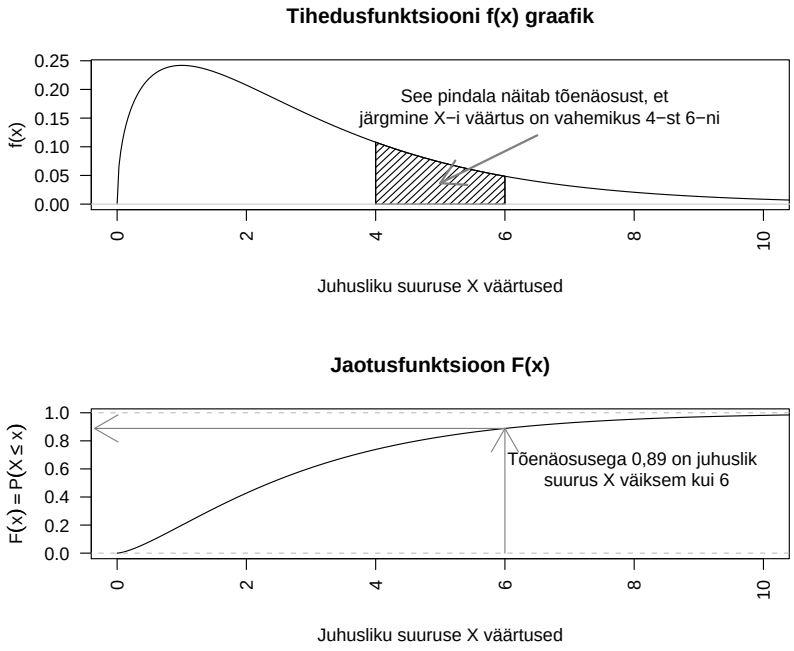
ja vastupidi — jaotusfunktsiooni teades saame leida tihedusfunktsiooni:

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

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.



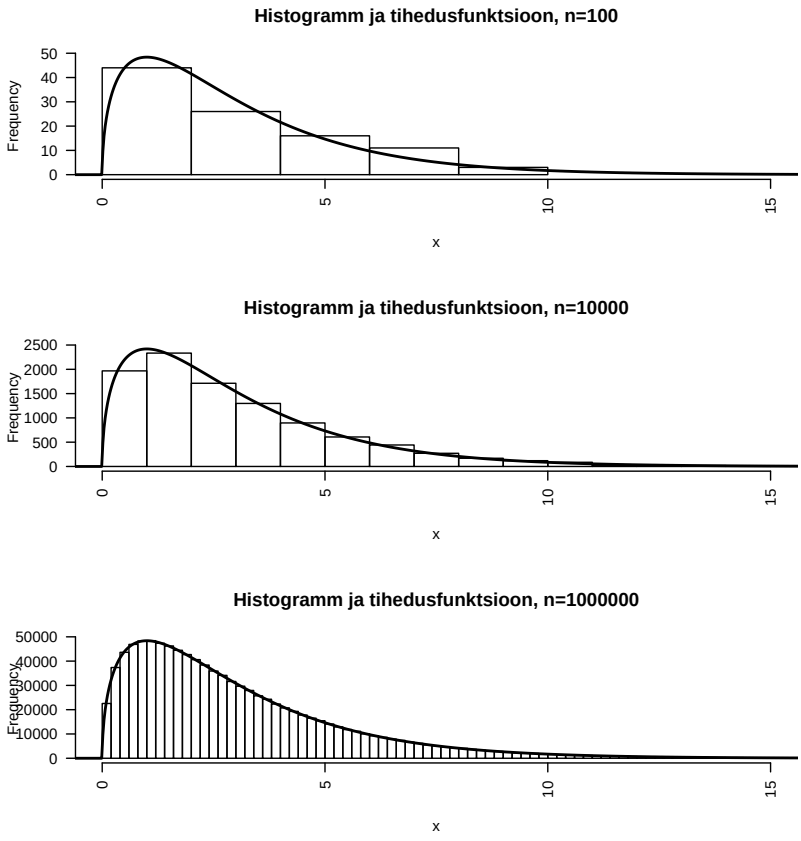
Joonis 9.3: Ühe pideva juhusliku suuruse tihedus- ja jaotusfunktsioon

Näide 9.5 *Joonistele 9.4 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 9.4: Tihedusfunktsioon ja histogram

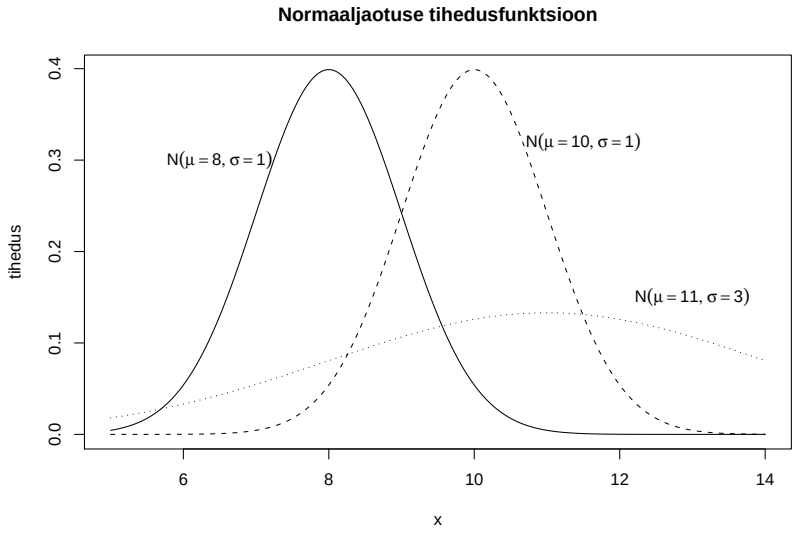
sioon on esitatav järgmisel kujul:

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

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 9.5.

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



Joonis 9.5: 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 $g'(\mu) \neq 0$),

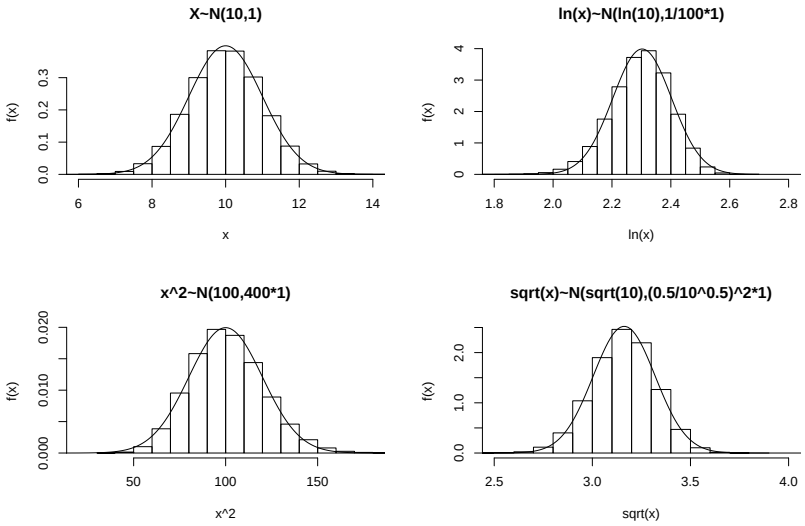
$$Y \sim N(g(X); \{g'(\mu)\}^2\sigma^2).$$

Seda fenomeeni iseloomustab ka joonis 9.6.

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 9.6: 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.