II

TÕENÄOSUSTEOORIA JA MATEMAATILINE STATISTIKA, TEOREETILISED JAOTUSED, PARAMEETRITE HINDAMINE

Antud peatükk püüab anda lühiülevaate matemaatilise statistika olemusest, tõenäosuse mõistest, juhuslikest suurustest, teoreetilistest jaotustest, jaotuste peamistest parameetritest ja nende hindamismeetoditest.

2.1 SÜNDMUS JA TÕENÄOSUS

2.1.1 Katse ja sündmus

Def. Katse (ekspermendi, vaatluse) tulemuseks on **sündmus**.

Näiteks on kahe heterosügootse indiviidi ristamine katse, samuti heterosügootse järglase saamine aga katse tulemus ehk sündmus.

Mündivise on katse, kulli või kirja pealetulek on sündmus.

Def. Üldisemalt on sündmus defineeritav kui **katsetulemuste hulk**.

Näiteks ristates kahte heterosügootset, genotüüpidega *Aa*, indiviidi, koosneb sündmus „järglane on homosügootne“ kahest katsetulemusest – sündmus „järglane on homosügootne“ leiab aset nii siis, kui järglane on genotüübiga *AA* kui ka siis, kui järglane on genotüübiga *aa*.

Kuuetahulise täringu viskamisel on katsetulemuseks 1, 2, 3, 4, 5 või 6 silma pealejäämine. Sündmus „tulemuseks on paarisarv silmi“ koosneb katsetulemustest 2, 4 ja 6.

Sündmusi tähistatakse tavaliselt ladina tähestiku suurtähtedega tähestiku algusest: A, B jne.

2.1.2 Tehted sündmustega ja sündmustevahelised seosed

Kuna sündmuste näol on tegu katsetulemuste hulkadega, on sündmustevahelised tehted analoogsed hulkadevaheliste tehetega.

Def. Olulisemad tehted on

- sündmuste A ja B **summa** $A \cup B$ (ehk $A + B$), mis toimub siis, kui toimub kas sündmus A, sündmus B või mõlemad;
- sündmuste A ja B **korrutis** $A \cap B$, mis toimub siis, kui toimuvad mõlemad sündmused A ja B;
- sündmuste A ja B **vahe** $A \setminus B$, mis toimub siis, kui toimub sündmus A, aga ei toimu sündmust B;
- sündmuse A **vastandsündmus** $\bar{A}$ (ehk A^c), mis toimub siis, kui ei toimu sündmust A.

Näiteks sündmus, et heterosügootne vanem genotüübiga *Aa* pärandab mõnele oma järglastest alleeli *A*, toimub nii siis, kui alleeli *A* saab päranduseks esimene järglane kui ka siis, kui alleeli *A* saab päranduseks teine või kolmas järglane või pärivad alleeli *A* kõik vaatlusaluse indiviidi järglased. St, et huvipakkuv sündmus on esitatav summana

„vanem pärandab mõnele oma järglastest alleeli *A*“ = „1. järglane pärib alleeli *A*“
∪ „2. järglane pärib alleeli *A*“ ∪ … ∪ „viimane järglane pärib alleeli *A*“.

Aga sündmus

„vanem pärandab kõigile oma järglastele alleeli *A*“ = „1. järglane pärib alleeli *A*“
∩ „2. järglane pärib alleeli *A*“ ∩ … ∩ „viimane järglane pärib alleeli *A*“

toimub vaid siis, kui kõik järglased on saanud päranduseks alleeli *A*.

Def. Kui mingi sündmuse jaoks ei leidu ühtki soodsat katsetulemust (sündmusele vastav katsetulemuste hulk on tühi), siis nimetatakse seda sündmust **võimatuks sündmuseks**. Võimatu sündmuse tähis on $\emptyset$.

Def. Kui mingi sündmuse jaoks on kõik katsetulemused soodsad, siis on see sündmus **kindel sündmus**. Kindla sündmuse tähis on Ω.

Def. Sündmust, mis pole ei kindel ega võimatu, nimetatakse **juhuslikuks sündmuseks**. Juhusliku sündmuse toimumine või mittetoimumine sõltub juhusest, st. sellest, missuguse tulemuseni katse sooritamisel jõuti.

Def. Kui sündmuse A jaoks soodsate katsetulemuste hulk sisaldub sündmuse B jaoks soodsate katsetulemuste hulgas, siis järeldub sündmuse A toimumisest sündmuse B toimumine. Seda sisaldussuhet märgitakse $A \subset B$.

Näiteks on päikese tõusmine hommikul kindel sündmus, 10 silma tulek kuuetahulise täringu viskamisel võimatu sündmus ning tuttava kohtamine tänava või võitmine loteriil juhuslikud sündmused.

Sündmus, et alleeli A suhtes homosügootne vanem pärandab oma järglasele alleeli A, on kindel sündmus;
sündmus, et see sama vanem pärandab järglasele alleeli a, on võimatu sündmus;
sündmus, et heterosügootne vanem genotüübiga Aa pärandab järglasele alleeli A, on juhuslik sündmus.

Sündmus „heterosügootne vanem genotüübiga Aa pärandab järglasele alleeli A“ sisaldub sündmuses „heterosügootsete vanemate paar $Aa \times Aa$ pärandab järglasele alleeli A“, sest esimese toimumisest järeldub ka teise toimumine.

Def. Sündmusi, mis ei saa üheaegselt toimuda, nimetatakse (üksteist) **välistavateks** sündmusteks.

Näiteks sündmused, et parajasti poegival „lehmal sünnib üks vasikas“ ja „lehmal sünnib kaks vasikat“ on üksteist välistavad sündmused, sest nad ei saa samaaegselt toimuda.

Omadused

1. Katsetulemused on alati üksteist välistavad (näiteks loom ei saa järglasele pärandada mõlemat oma alleeli või lehm ei saa korraga anda mitmes erinevas koguses piima).
2. Sündmus ja tema vastandsündmus on üksteist välistavad.
3. Kui sündmused on üksteist välistavad, siis on nende korrutis võimatu, $A \cap B = \emptyset$.

2.1.3 Sündmuse tõenäosus

Def. Sündmuse A **tõenäosus** $P(A)$ on selle sündmuse jaoks soodsate katsetulemuste arvu k ja kõigi katsetulemuste arvu n suhe:

$$P(A) = {}^{n}\!/_{k}\,. \tag{2.1}$$

Näiteks on tõenäosus, et heterosügootne vanem genotüübiga Aa pärandab järglasele alleeli A, ½:

$$P(Aa \rightarrow A) = ½.$$

Sest kõigi katsetulemuste arv (pärandatavate geenivariantide arv, A või a) on kaks, millest vaid üks (alleeli A pärandumine) on huvipakkuva sündmuse jaoks soodne variant.

Omadused

1. Tõenäosus on arv 0 ja 1 vahel.
2. Mida suurem on sündmuse tõenäosus, seda rohkem on alust loota selle sündmuse toimumist.

3. Kui sündmuste A ja B vahel on (range) sisaldusseos: $A \subset B$, siis kehtib võrratus $P(A) < P(B)$.
4. Võimatu sündmuse tõenäosus on null: $P(\varnothing) = 0$.
5. Kindla sündmuse tõenäosus on üks: $P(\Omega) = 1$.

Näiteks sellest, et sündmus „heterosügootne vanem genotüübiga *Aa* pärandab järglasele alleeli *A*“ sisaldub sündmuses „heterosügootsete vanemate paar *Aa*×*Aa* pärandab järglasele alleeli *A*“, järeldub vastavalt 3. omadusele, et

P(heterosügootne vanem genotüübiga *Aa* pärandab järglasele alleeli *A*)
< P(heterosügootsete vanemate paar *Aa*×*Aa* pärandab järglasele alleeli *A*),

mis, kui järele mõelda, on ju loomulik.

2.1.4 Tehted tõenäosustega

Def. Kahe teineteist **välistava** sündmuse A ja B **summa tõenäosus** võrdub liidetavate tõenäosuste summaga:

$$P(A \cup B) = P(A) + P(B), \text{ kui } A \cap B = \varnothing. \tag{2.2}$$

Näiteks sündmused „lehmal sünnib üks vasikas“ ja „lehmal sünnib kaks vasikat“ on teineteist välistavad, mistap avaldub sündmuse „lehmal sünnib vähemalt kaks vasikat“ tõenäosus kujul

P(lehmal sünnib vähemalt kaks vasikat)
= P(lehmal sünnib üks vasikas) + P(lehmal sünnib kaks vasikat).

Def. Üldjuhul, kui sündmused A ja B ei ole üksteist välistavad, avaldub nende **summa tõenäosus** kujul

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{2.3}$$

Def. Sündmuse A **vastandsündmuse** $\bar{A}$ **tõenäosus** avaldub sündmuse A tõenäosuse kaudu kujul:

$$P(\bar{A}) = 1 - P(A). \tag{2.4}$$

Näiteks dialleelsel juhul avaldub alleeli *a* esinemissagedus (tõenäosus) populatsioonis $P(a)$ alleeli A esinemissageduse (tõenäosuse) $P(A)$ kaudu kujul

$$P(a) = 1 - P(A).$$

Def. Kui ühe sündmuse toimumine ei mõjuta teise sündmuse toimumist, siis on need sündmused **sõltumatud**.

Sõltumatute sündmuste A ja B **korrutise tõenäosus** võrdub sündmuste A ja B tõenäosuste korrutisega:

$$P(A \cap B) = P(A)P(B). \tag{2.5}$$

Viimast võrdust kasutatakse ka defineerimaks sündmuste sõltumatust: **kui sündmuste korrutise tõenäosus võrdub sündmuste tõenäosuste korrutisega, siis on need sündmused sõltumatud**.

Näide 1. See, millise alleeli pärandab vanem ühele järglasele, on sõltumatu sellest, millise alleeli pärib teine järglane, siis avaldub tõenäosus, et vanem pärandab kõigile oma järglastele sama alleeli, näiteks *A*, kujul

P(vanem pärandab kõigile oma järglastele alleeli *A*) = P(1. järglane pärib alleeli *A*)
× P(2. järglane pärib alleeli *A*) × … × P(viimane järglane pärib alleeli *A*).

Juhul, kui vanem oli homosügootne genotüübiga *AA*, on üleltoodud tõenäosus üks, sest alleeli *A* pärandamine on kindel sündmus kõigi järglaste puhul, kindla sündmuse tõenäosus on üks, mistap võrdub ühega ka nende sündmuste tõenäosuste korrutis.

Kui aga vanem on heterosügootne genotüübiga *Aa*, siis P(järglane pärib alleeli *A*) = ½, ja seda iga järglase puhul. Seega

P(vanem pärandab kõigile oma järglastele alleeli A) = ½ × ½ × … × ½ = $(½)^n$,

kus n on järglaste arv.

Näide 2. Mendeli II seadus e alleelide lahknemise seadus postuleerib teatavasti, et heterosügootsete indiviidide omavahelisel ristamisel toimub järglaspõlvkonnas tunnuste lahknemine kindlates sagedussuhetes (nn lahknemissuhetes). Näiteks vanempaari Aa×Aa järglasest on ¼ genotüübiga AA, ½ genotüübiga Aa ja ¼ genotüübiga aa. Toodud genotüüpide sagedused tulenevad aga otseselt elementaarsest tõenäosusteooriast.

Tõenäosus, et heterosügootse vanempaari Aa×Aa järglane on genotüübiga AA avaldub kujul

P(järglase genotüüp on AA) = P(emalt pärandub alleel A ja isalt pärandub alleel A) = …

(et see, milline alleel pärandub isalt, on sõltumatu sellest, milline alleel pärandus emalt, on sündmuste korrutise tõenäosus viimases avaldises esitatav sündmuste tõenäosuste korrutisena)

… = P(emalt pärandub alleel A)×P(isalt pärandub alleel A) = ½ × ½ = ¼.

Tõenäosus, et heterosügootse vanempaari Aa×Aa järglane on genotüübiga Aa, avaldub kujul

P(järglase genotüüp on Aa)
= P[(emalt pärandub alleel A ja isalt pärandub alleel a)
või (emalt pärandub alleel a ja isalt pärandub alleel A)] = …

(sündmused „emalt pärandub alleel A ja isalt pärandub alleel a“ ning „emalt pärandub alleel a ja isalt pärandub alleel A“ on üksteist välistavad – nad ei saa toimuda samaaegselt – mistap võrdub nende sündmuste summa tõenäosus tõenäosuste summaga)

… = P(emalt pärandub alleel A ja isalt pärandub alleel a)
+ P(emalt pärandub alleel a ja isalt pärandub alleel A) = …

(viimases avaldises kujutavad mõlemad tõenäosused enesest sõltumatute sündmuste korrutise tõenäosusi, mistap saab need avaldada tõenäosuste korrutistena)

… = P(emalt pärandub alleel A)×P(isalt pärandub alleel a)
+ P(emalt pärandub alleel a)×P(isalt pärandub alleel A) = ½×½ + ½×½ = ¼ + ¼ = ½.

Tõenäosus, et järglase genotüüp on aa, avaldub analoogselt järglase genotüübi AA tõenäosusega.

2.1.5 Tinglik tõenäosus

Def. Sündmuse A **tinglikuks tõenäosuseks** sündmuse B suhtes nimetatakse sündmuse A tõenäosust eeldusel, et toimus sündmus B.

Sündmuse A tinglikku tõenäosust sündmuse B suhtes tähistatakse P(A|B).

Omadused

1. Kui sündmused A ja B on sõltumatud, siis P(A|B) = P(A) ja P(B|A) = P(B), ehk sündmuse A toimumine ei mõjuta sündmuse B toimumist ja vastupidi.
2. Kui sündmused A ja B on üksteist välistavad, siis P(A|B) = 0 ja P(B|A) = 0, sest eeldusel, et üks sündmus on juba toimunud, on teise sündmuse toimumine võimatu ja selle tõenäosus null.

Def. Sündmuse A **tinglik tõenäosus** sündmuse B suhtes avaldub valemiga

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}. \tag{2.6}$$

Viimasest avaldisest järeldub ka **sündmuste korrutise tõenäosus** üldjuhul:

$$\mathrm{P}(A \cap B) = \mathrm{P}(A \mid B)\,\mathrm{P}(B). \tag{2.7}$$

Näide. Oletame, et genotüübiga *AA* ja *Aa* indiviidid on kollased ning genotüübiga *aa* indiviidid rohelised.

Vanempaari *Aa*×*Aa* järglasest on ¼ genotüübiga *AA*, ½ genotüübiga *Aa* ja ¼ genotüübiga *aa*. Kollaseid järglaseid on $P(AA \cup Aa) = P(AA) + P(Aa) = ¼ + ½ = ¾$ ja rohelisi järglaseid on $P(aa) = ¼$.

Kui suur on tõenäosus, et juhuslikult valitud kollane järglane on genotüübiga *Aa*?

P(genotüüp = *Aa* | järglane on kollane)
= P(genotüüp = *Aa* ∩ järglane on kollane) / P(järglane on kollane) = ½ / ¾ = ⅔.

2.1.6 Täistõenäosuse valem

Sageli peavad sündmuse *A* toimumiseks olema eelnevalt toimunud mingid teised sündmused H_i, $i = 1, \ldots, n$. Tõenäosusteoorias nimetatakse neid sündmuse *A* toimumisele eelnevaid sündmusi hüpoteesideks.

Kui nüüd eeldada, et

- sündmus *A* saab toimuda vaid mingi sündmuse H_i toimumise järgselt,
- sündmuste loetelu $H_1, \ldots, H_n$ sisaldab kõiki võimalikke variante sündmuse *A* toimumisele eelnevatest sündmustest ja
- sündmused $H_1, \ldots, H_n$ on üksteist välistavad,

siis on sündmuse *A* toimumise tõenäosus (nn **täistõenäosus**) leitav valemist

$$P(A) = P(H_1)P(A|H_1) + P(H_2)P(A|H_2) + \ldots + P(H_n)P(A|H_n). \quad (2.8)$$

Näide. Talumehel oli kolm jäära. Paaritusperioodil lasi ta neist esimese utede hulka üheks, teise kaheks ja kolmanda kolmeks päevaks. Kõik jäärad olid genotüpiseeritud tallede kasvukiirust potentsiaalselt mõjutava dialleelse lookuse osas, misläbi oli teada, et esimene ja teine jäär olid kasvukiirust pärssiva retsessiivse alleeli kandjad, kolmas jäär oli aga homosügootne kasvukiirust suurendava dominantse alleeli suhtes. Eeldades, et paaritatud uttede arv on proportsionaalne karjas viibitud ajaga ja et mitmike sünni sagedus on samasugune kõigi jäärade korral, saab leida, kui suure tõenäosusega on suvaline viie kuu pärast sündiv tall pärinud isalt kasvukiirust pärssiva alleeli.

Vastavalt täistõenäosuse valemile saab otsitava tõenäosuse esitada kujul:

P(isalt on pärandunud kasvukiirust pärssiv alleel)
= P(isaks on esimene jäär) × P(esimeselt jääralt pärandus kasvukiirust pärssiv alleel)
+ P(isaks on teine jäär) × P(teiselt jääralt pärandus kasvukiirust pärssiv alleel)
+ P(isaks on kolmas jäär) × P(kolmandalt jääralt pärandus kasvukiirust pärssiv alleel)
= (⅙ × ½) + (⅓ × ½) + (½ × 0) = 3/12 = ¼ = 0,25.

Ehk siis hoolimata sellest, et halva alleelita jäär oli karjas sama kaua, kui teised kokku, ja teiste jäärade puhul oli halva alleeli pärandumise tõenäosus vaid ½, pärivad keskmiselt 25% järglastest isa poolt siiski mittesoovitava kasvukiirust pärssiva retsessiivse alleeli.

2.1 JUHUSLIKUD SUURUSED JA OLULISEMAD NEID ISELOOMUSTAVAD PARAMEETRID

2.2.1 Juhuslik suurus

Sarnaselt eelmises alapeatükis käsitletud sündmuse mõistele on ka juhuslik suurus defineeritud juhusliku katse kaudu. Kui sündmus kujutab enesest katsetulemust või katsetulemuste hulka, siis **juhuslik suurus** on katsetulemuse arvuline (kvantitatiivne) resultaat. Et katse all võib mõista mistahes eksperimenti, mõõtmist või vaatlust, on juhuslikeks suurusteks ka andmeanalüüsil analüüsitavad tunnused (lehmade piimatoodang, emiste pesakonna suurus jmt), samuti kõikvõimalikud hinnangulised näitajad ja teoreetiliste mudelite liikmed (aretusväärtus, söötmise mõju, genotüübiefekt, keskkonnaefekt jmt). Tunnuste mõõtmistulemused konkreetsetel loomadel (valim) või tegelikud aretusväärtused või konkreetsete geenikombinatsioonide mõjud kujutavad enesest **juhuslike suuruste realiseerunud väärtuseid**.

Juhuslikke suurusi tähistatakse traditsiooniliselt suurte tähtedega X, Y, Z jmt ning nende väärtuseid vastavate väiketähtedega x, y, z jne. Vajadusel lisatakse tähistusele indeksid.

2.2.2 Keskväärtus

Def. Juhusliku suuruse **keskväärtuseks** (ooteväärtuseks e oodatavaks väärtuseks, inglise keeles *expected value*, *expectation*) nimetatakse tema lõpmatu hulga väärtuste keskmist. Juhusliku suuruse X keskväärtust tähistatakse $\mathrm{E}(X)$, sageli kasutatakse ka tähte μ.

Omadused

1. $\min(X) \le \mathrm{E}(X) \le \max(X)$.
2. $\mathrm{E}(k) = k$, k on konstant.
3. $\mathrm{E}(X + k) = \mathrm{E}(X) + k$.
4. $\mathrm{E}(kX) = k \times \mathrm{E}(X)$.
5. $\mathrm{E}(X_1 + X_2) = \mathrm{E}(X_1) + \mathrm{E}(X_2)$.
6. Kui X_1 ja X_2 on sõltumatud, siis $\mathrm{E}(X_1X_2) = \mathrm{E}(X_1)\mathrm{E}(X_2)$.
7. Kui $\mathbf{X} = (X_1 \; X_2 \; \ldots \; X_n)^{\mathrm{T}}$, siis $\mathrm{E}(\mathbf{X}) = \mathrm{E}(X_1) \; \mathrm{E}(X_2) \; \ldots \; \mathrm{E}(X_n)^{\mathrm{T}}$.

2.2.2 Dispersioon

Def. Juhusliku suuruse X **dispersioon** $\mathrm{var}(X)$ (ka $\mathrm{V}(X)$, $\mathrm{D}(X)$, σ_X^2, inglise keeles *variance*) on defineeritud seosega

$$\mathrm{var}(X) = \mathrm{E}[X - \mathrm{E}(X)]^2 = \mathrm{E}(X^2) - \mathrm{E}(X)\mathrm{E}(X). \tag{2.9}$$

Omadused

1. $\mathrm{var}(X) \ge 0$.
2. $\mathrm{var}(k) = 0$, k on konstant.
3. $\mathrm{var}(X + k) = \mathrm{var}(X)$.
4. $\mathrm{var}(kX) = k^2\mathrm{var}(X)$.
5. Kui X_1 ja X_2 on sõltumatud, siis $\mathrm{var}(X_1 + X_2) = \mathrm{var}(X_1) + \mathrm{var}(X_2)$.

Ruutjuurt juhusliku suuruse dispersioonist nimetatakse **standardhälbeks** – $\sqrt{\mathrm{var}(X)} = \sigma_X$.

2.2.3 Kovariatsioon

Def. Juhuslike suuruste X_1 ja X_2 vaheline **kovariatsioon** $\mathrm{cov}(X_1, X_2)$ (ka $\sigma_{X_1X_2}$, inglise keeles *covariance*) kirjeldab nende juhuslike suuruste vahelist lineaarset statistilist sõltuvust ja on defineeritud seosega

$$\mathrm{cov}(X_1, X_2) = \mathrm{E}(X_1X_2) - \mathrm{E}(X_1)\,\mathrm{E}(X_2). \tag{2.10}$$

Omadused

1. $\mathrm{cov}(X_1, X_2) = \mathrm{cov}(X_2, X_1)$.
2. $\mathrm{cov}(X, X) = \mathrm{var}(X)$.
3. $\mathrm{cov}(aX_1, bX_2) = ab\,\mathrm{cov}(X_1, X_2)$, a, b on konstandid.
4. $\mathrm{var}(X_1 \pm X_2) = \mathrm{var}(X_1) + \mathrm{var}(X_2) \pm 2\,\mathrm{cov}(X_1, X_2)$.
5. Kui X_1 ja X_2 on sõltumatud, siis $\mathrm{cov}(X_1, X_2) = 0$ (sellest ja eelmisest omadusest järeldub dispersiooni 3. omadus).

Def. Enam kui kahe juhusliku suuruse korral on nende dispersioone ja kovariatsioone mugav esitada maatriksina: juhuslike suuruste vektori $\mathbf{X} = (X_1\ X_2\ \ldots\ X_n)^{\mathrm{T}}$ **dispersioonimaatriksiks** (kovariatsioonimaatriksiks) nimetatakse maatriksit

$$\begin{aligned}\mathrm{var}(\mathbf{X}) &= \mathrm{E}(\mathbf{X} - \mathrm{E}\,\mathbf{X})(\mathbf{X} - \mathrm{E}\,\mathbf{X})^{\mathrm{T}} \\ &= \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_n) \\ \mathrm{cov}(X_1, X_2) & \mathrm{var}(X_2) & \ldots & \mathrm{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_1, X_n) & \mathrm{cov}(X_2, X_n) & \ldots & \mathrm{var}(X_n) \end{pmatrix} = \begin{pmatrix} \sigma^2_{X_1} & \sigma_{X_1X_2} & \cdots & \sigma_{X_1X_n} \\ \sigma_{X_1X_2} & \sigma^2_{X_2} & \cdots & \sigma_{X_2X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_1X_n} & \sigma_{X_2X_n} & \cdots & \sigma^2_{X_n} \end{pmatrix},\end{aligned} \tag{2.11}$$

kus peadiagonaalil paiknevad juhuslike suuruste dispersioonid ja väljaspool peadiagonaali juhuslike suuruste vahelised kovariatsioonid.

Dispersioonimaatriksi omadused on analoogsed kovariatsiooni ja dispersiooni omadustega, väikesed erinevused tulenevad vaid sellest, et maatriksite korrutamisel ei kehti kõik üksikelementide korrutamise reeglid. Näiteks dispersioon konstantse maatriksiga **A** korrutatud juhuslike suuruste vektorist **X** avaldub kujul $\mathrm{var}(\mathbf{AX}) = \mathbf{A}\,\mathrm{var}(\mathbf{X})\mathbf{A}^{\mathrm{T}}$.

2.2.4 Korrelatsioon

Def. Jagades juhuslike suuruste vahelise kovariatsiooni läbi nende standardhälvete korrutisega saame normeeritud kovariatsiooni, mida nimetatakse **korrelatsioonikordajaks**:

$$\mathrm{r}(X_1, X_2) = \frac{\mathrm{cov}(X_1, X_2)}{\sqrt{\mathrm{var}(X_1)\,\mathrm{var}(X_2)}}. \tag{2.12}$$

Omadused

1. $\mathrm{r}(X_1, X_2) = \mathrm{r}(X_2, X_1)$.
2. $\mathrm{r}(X, X) = 1$.
3. Kui X_1 ja X_2 on sõltumatud, siis $\mathrm{r}(X_1, X_2) = 0$.
4. $|\mathrm{r}(X_1, X_2)| \le 1$.

Sarnaselt kovariatsioonimaatriksiga defineeritakse juhuslike suuruste vektori $\mathbf{X} = X_1\ X_2\ \ldots\ X_n^{\mathrm{T}}$ **korrelatsioonimaatriks** $\mathrm{r}(\mathbf{X})$ elementidega r_{ij}, $i, j = 1,\ldots,n$, kus

$$r_{ij} = r(X_i, X_j) = \mathrm{cov}(X_i, X_j) \big/ \sqrt{\mathrm{var}(X_i)\,\mathrm{var}(X_j)} \quad \text{ja} \quad r_{ii} = r(X_i, X_i) = 1.$$

Nii kovariatsiooni- kui ka korrelatsioonimaatriksid on sümmeetrilised maatriksid.

2.2.5 Lineaarne regressioon

Prognoosimaks juhusliku suuruse Y käitumist tingimusel, et sellega lineaarses sõltuvuses olev teine juhuslik suurus X omandab mingi kindla väärtuse, kasutatakse lineaarset regressiooniseost kujul

$$\mathrm{E}(Y) = a + bX \ .$$

Võrrandi vasakul poolel olev suurus $\mathrm{E}(Y)$ näitab, et tulemuseks on juhusliku suuruse Y oodatav (keskmine) väärtus juhusliku suuruse X mingi väärtuse korral.

Regressioonivõrrandi kordajad a ja b hinnatakse seostest

$$\hat{b} = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)} \quad \text{ja} \quad \hat{a} = \mathrm{E}\,Y - \hat{b}\,\mathrm{E}\,X \ . \tag{2.13}$$

Kasuks tuleb teadmine, et kordaja b väljendab juhusliku suuruse Y muutumise suurust juhusliku suuruse X muutumisel ühe ühiku võrra.

2.3 TEOREETILISED JAOTUSED

Juhuslike suuruste väärtusi ja nende paiknemist iseloomustatakse **jaotusseadustega**. Viimased kujutavad enesest parameetritest sõltuvaid matemaatilisi eeskirju, nn **teoreetilisi jaotusi**, mille alusel on võimalik tuvastada juhusliku suuruse väärtuste hulk ja leida väärtuste esinemise tõenäosused. Konkreetse juhusliku suuruse iseloomustamiseks sobiv jaotusseadus on määratud juhusliku suuruse tekkemehhanismi e olemuse läbi. Teoreetilised jaotused, mida on kahte tüüpi – diskretsed ja pidevad jaotused –, on aluseks statistiliste hüpoteeside kontrollimisel, sageli ka parameetrite väärtuste ja nende hinnangute usaldusväärsuse hindamisel.

Def. Diskreetne jaotus esitatakse tõenäosusfunktsiooniga

$$p(k) = \mathrm{P}(X = k)$$

või jaotustabeliga $\{p(k), k\}$, mis defineerib täpselt ära iga üksiku väärtuse esinemise tõenäosuse (k on juhusliku suuruse võimalik väärtus).

Def. Pidev jaotus esitatakse tihedusfunktsiooniga

$$f(x) = \mathrm{d}F(x)/\mathrm{d}x,$$

mille abil on defineeritud juhusliku suuruse iga väärtuse mingisse fikseeritud vahemikku (a, b) sattumise tõenäosus:

$$P(a < X < b) = \int_a^b f(x)dx = F(b) - F(a),$$

$F(x) = \mathrm{P}(X \le x)$ on jaotusfunktsiooni väärtus kohal x.

Diskreetsed jaotused, mis tekivad millegi kokkulugemisel, subjektiivsel hindamisel või mingi nähtuse toimumise või mittetoimumise fikseerimisel, vajavad analüüsil siinses kursuses käsitletavate meetodite edasiarendusi – nn üldistatud lineaarseid mudeleid – mistõttu on järgnevalt tutvustatud vaid traditsioonilises üldiste lineaarsete mudelite teoorias ja geneetiliste parameetrite hindamisel kasutust leidvaid pidevaid jaotusi (ausalt öeldes neid ehk selle lühikursuse raames vaja ei lähegi, aga kui tekib soov või vajadus peale aretusväärtuste hindamise tutvuda ka dispersioonikomponentide ja päritavuskoefitsientide hindamisega, on vaja teadmisi normaaljaotusest, ning kui huvi pakub hinnangute täpsus ja hüpoteeside kontrollimine mingi parameetri nullist erinemise kohta, kulub ära ettekujutus χ^2-, F- ja t-jaotusest).

2.3.1 Normaaljaotus

Def. Normaaljaotusega juhuslikku suurust X keskväärtusega $\mathrm{E}(X) = \mu$ ja dispersiooniga $\mathrm{var}(X) = \sigma^2$ tähistatakse $X \sim N(\mu, \sigma^2)$ ja tema tihedusfunktsioon esitub valemiga

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}. \qquad (2.14)$$

Keskväärtuse ja dispersiooni, kui normaaljaotuse ainukeste parameetrite läbi, on tihedusfunktsioon iga juhusliku suuruse X väärtuse tarvis määratud.

Tähtsaks omaduseks on, et normaaljaotusega juhuslike suuruste lineaarkombinatsioon on samuti normaaljaotusega (muutuvad vaid parameetrite väärtused). Sagedaseimaks lineaarteisenduseks on standardiseerimine

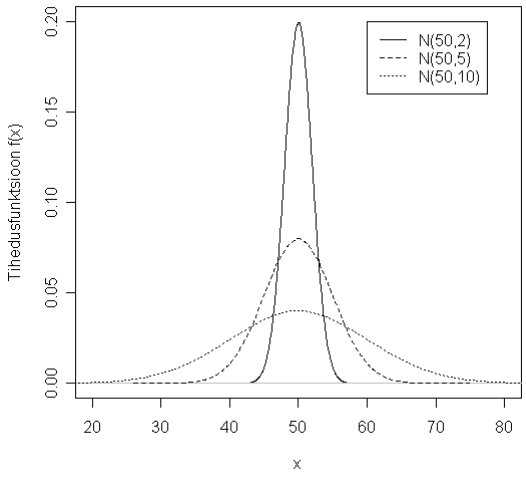
$$Z = \frac{X - \mu}{\sigma} \sim N(0,1),$$

kus $N(0,1)$ on standardne normaaljaotus, mille jaotusfunktsiooni $\Phi(x)$ väärtused on tabuleeritud (vt tabel 2.1). Seejuures kehtivad seosed $F(x) = \Phi(\frac{x-\mu}{\sigma})$ ja $\Phi(-x) = 1 - \Phi(x)$.

Tabel 2.1. Standardse normaaljaotuse enamkasutatavad jaotusfunktsiooni väärtused; $\Phi(x) = P(X \le x)$, kus $X \sim N(0,1)$.

$\Phi(x)$	0,005	0,025	0,05	0,5	0,95	0,975	0,995
x	-2,58	-1,96	-1,64	0	1,64	1,96	2,58

Näide 1. Normaaljaotusega $N(50,\sigma^2)$ on näiteks vere kogus indiviidi 50 ml vereproovis, kus σ^2 iseloomustab proovivõtmise täpsust.



Joonis 2.1. Näiteid normaaljaotuse $N(50,\sigma)$ tihedusfunktsiooni graafikuist erinevate parameetri σ väärtuste korral.

Näide 2. Piimaveiste suhteline piimajõudluse aretusväärtus SPAV väljendatakse punktides keskväärtusega 100 ja standardhälbega 12 punkti. Kuna aretusväärtuste hindamisel eeldatakse nende normaaljaotust (sellest tuleb täpsemalt juttu edasistes loengutes), võib arvata, et ka SPAV-i väärtused jaotuvad normaaljatuse järgi, ehk täpsemalt SPAV ~ N(100, 12).

Tehtud eelduse alusel võib näiteks leida, milline on see vahemik, kuhu jääb 95% piimaveiste suhteline piimajõudluse aretusväärtus SPAV.

Vastavalt standardse normaaljaotuse jaotusfunktsiooni väärtustele

$$\mathrm{P}[(\mathrm{SPAV}-100)/12 < -1{,}96] = 0{,}025 \quad \text{ja} \quad \mathrm{P}[(\mathrm{SPAV}-100)/12 < 1{,}96] = 0{,}975,$$

kus (SPAV–100)/12 ~ N(0, 1).

Siit edasi on juba loogilne, et

$$\mathrm{P}[-1{,}96 < (\mathrm{SPAV}-100)/12 < 1{,}96] = 0{,}95,$$

millest

$$\mathrm{P}(100 - 1{,}96\times 12 < \mathrm{SPAV} < 100 + 1{,}96\times 12) = \mathrm{P}(76{,}5 < \mathrm{SPAV} < 123{,}5) = 0{,}95.$$

Seega jääb 95% piimaveiste SPAV vahemikku 76,5-123,5 punkti.

Samuti saab näiteks leida, kui suur on üle 120-punktilise SPAV-ga veiste hulk.

$$\mathrm{P}(\mathrm{SPAV} > 120) = 1 - \mathrm{P}(\mathrm{SPAV} < 120) = 1 - \mathrm{P}[(\mathrm{SPAV}-100)/12 < (120-100)/12]$$

$$= 1 - \Phi(1{,}667) = 1 - 0{,}952 = 0{,}048.$$

Seega peaks enam kui 120-punktlise SPAV-ga olema eeldatavalt 4,8% Eesti piimaveiseid.

2.3.2 χ^2-jaotus

Def. Sõltumatute standardse normaaljaotusega juhuslike suuruste $X_1,\ldots,X_n$ ruutude summa on χ^2-jaotusega vabadusastmete arvuga n. Sümbolite kaudu defineeritult: kui $X_1,\ldots,X_n$ on sõltumatud juhuslikud suurused, kus $X_i \sim N(0,1)$, $i = 1,\ldots,n$, siis

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n) \quad \text{ja} \quad \sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \chi^2(n-1),$$

kus $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Kui $X \sim \chi^2(n)$, siis $\mathrm{E}(X) = n$ ja $\mathrm{var}(X) = 2n$.

2.3.3 *F*-jaotus

Def. Kui juhuslikud suurused $U \sim \chi^2(k_1)$ ja $V \sim \chi^2(k_2)$ ning U ja V on sõltumatud, siis on juhuslik suurus Z F-jaotusega:

$$Z = \frac{U/k_1}{V/k_2} \sim F(k_1, k_2).$$

Jaotuse parameetrid k_1 ja k_2 on positiivsed täisarvud, mida nimetatakse F-jaotuse vabadusastmeteks (inglise keeles *degrees of freedom*).

2.3.4 *Student*'i *t*-jaotus

Def. Kui juhuslik suurus $X \sim N(0,1)$ ja juhuslik suurus $Y \sim \chi^2(n)$, kusjuures X ja Y on sõltumatud, siis

$$Z = \frac{X}{\sqrt{Y/n}} \sim t(n),$$

ehk Z on t-jaotusega vabadusastmete arvuga n.

Matemaatilise statistika rakenduste tarvis on oluline tulemus, et kui $X_i \sim N(\mu, \sigma^2)$, $i = 1,...,n$, siis

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(n-1),$$

kus $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ja $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$.

2.4 PARAMEETRITE HINDAMINE

Kui me ka teame, mis teoreetilisele jaotusele meie poolt uuritav suurus oma olemuselt vastab, ei tea me siiski selle jaotuse parameetreid. Viimased tuleb hinnata tuginedes andmetele (valimile). Vastav arvutusvalem on lihtsamal juhul leitav intuitiivselt (nö mõistlikkuse printsiibist lähtudes – näiteks on loomulik võtta keskväärtuse hinnanguks valimi aritmeetiline keskmine). Keerulisemal juhul ei pruugi aga intuitsioonist piisata või peab seda intuitsiooni vähemalt kontrollima. Matemaatiline statistika pakub mitmeid võimalusi parameetrite hindamisvalemite teoreetiliseks konstrueerimiseks ning saadavate hinnangute headuse üle otsustamiseks. Levinuimad meetodid on suurima tõepära meetod ja vähimruutude meetod.

2.4.1 Suurima tõepära meetod (*maximum likelihood method*, ML-meetod)

Suurima tõepära meetodit kasutatakse siis, kui teoreetiline jaotus on teada ja hinnatav parameeter kujutab enesest selle teoreetilise jaotuse tihedus- või tõenäosusfunktsiooni parameetrit (argumenti). Hinnanguks on siis loomulik valida see parameetri väärtus, mis realiseerunud juhul (st uuritavate andmete korral) kõige paremini sobib ehk teisisõnu on antud valimi jaoks tõepäraseim väärtus. Et eelduse kohaselt sõltub hinnatavast parameetrist ka üldkogumi jaotus, siis on hinnangule vastav jaotus tõepäraseim antud valimi jaoks.

Teoreetilise definitsioonina sõnastades: parameetri θ suurima tõepära hinnanguks nimetatakse väärtust $\hat{\theta}$, mille korral tõepärafunktsioon $L(\theta)$ saavutab maksimaalse väärtuse oma parameeterruumis,

$$L(\theta) = \begin{cases} f(x_1;\theta)\cdot f(x_2;\theta)\cdot\ldots\cdot f(x_n;\theta), \text{ pideval juhul,} \\ p(x_1;\theta)\cdot p(x_2;\theta)\cdot\ldots\cdot p(x_n;\theta), \text{ diskreetsel juhul.} \end{cases}$$

Tõepärafunktsioon kujutab enesest sama valemit, mis tihedusfunktsioongi. Erinevus seisneb selles, et tõepärafunktsioonis loetakse vastupidiselt teoreetilistele jaotustele fikseerituks andmete osa (meil on ju olemas mingid realiseerunud väärtused) ja juhuslikuks, kirjeldamist vajavaks, parameetrite osa. Juhul, kui tihedus- või tõenäosusfunktsiooni abil defineeritud teoreetiline jaotus vastab tegelikkusele, on parameetrite suurima tõepära hinnangud täpseimad.

Näide. Olgu meil vaatluse all dialleelne lookus alleelide *a* ja *A* sagedustega populatsioonis vastavalt *p* ja 1 – *p*. Ja olgu eelnevalt ka teada, et alleeli *a* sagedus saab olla kas ½ või ¼, st $p\in\{½;¼\}$. Olgu meil kaks vaatlust: x_1=‘*a*’ ja x_2=‘*a*’. Kumb on tõepärasem hinnang *p*-le, kas ½ või ¼?

Tõepärafunktsioon: $L(p) = \mathrm{P}(X=x_1)\times\mathrm{P}(X=x_2) = p^2$, millest $L(½) = ¼$, $L(¼) = {}^1\!/_{16}$.

Kuna $L(½) > L(¼)$, siis $\hat{p} = ½$ on suurima tõepära hinnang *p*-le.

2.4.2 Vähimruutude meetod (*least square method*, LS-meetod)

Vähimruutude meetod ei eelda mingi tihedus- või tõenäosusfunktsiooni kasutamist, mistõttu on selle abil saadavad hinnangud sageli lihtsamal kujul võrreldes teiste hindamismeetoditega. Nagu suurima tõepära meetod, püüab vähimruutude meetodki valida hinnanguks seda parameetri väärtust, mis realiseerunud juhul (uuritavate andmete korral) kõige paremini sobib. Ainult sobivus on defineeritud pisut teisel kujul – parim hinnang on selline, mille korral ruuterinevus realiseerunud väärtuste ja parameetri hinnangule vastavate väärtuste vahel on minimaalne.

Näiteks lineaarse regressioonivõrrandi $\mathrm{E}(Y) = a + bX$ parameetrite hinnangufunktsioonid (2.13) on tuletatud just vähimruutude meetodil, minimiseerides avaldise

$$\mathrm{E}[Y - (a + bX)]^2 \to \min$$

a ja *b* suhtes.

2.4.3 Hinnangute omadused

Def. Parameetri θ hinnangut nimetatakse **nihketa hinnanguks** (*unbiased*), kui $\mathrm{E}(\hat{\theta}) = \theta$; ehk hinnang on „keskmiselt õige“, puudub süstemaatiline viga.

Näiteks on valimi keskmine nihketa hinnanguks populatsiooni keskväärtusele – valimi keskmine võib populatsiooni keskväärtusest olla samavõrra väiksem kui suurem, olles seega keskmiselt õige.

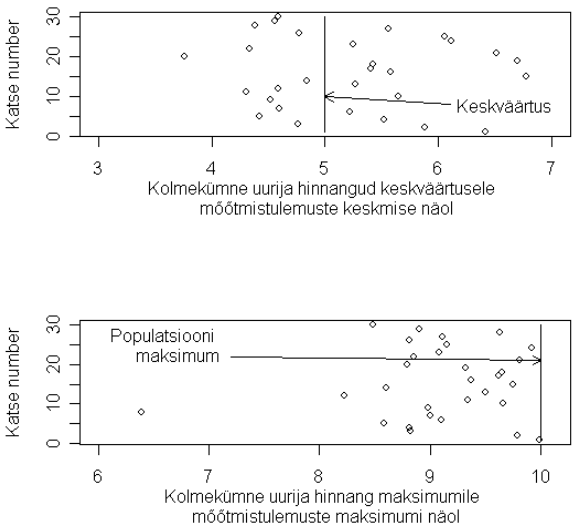
Matemaatiliselt tõestab väite, et valimi (juhusliku suuruse realiseerunud väärtuste) keskmine $\bar{x} = \frac{1}{n}\sum x_i$ on nihketa hinnang populatsiooni (juhusliku suuruse) keskväärtusele E(*X*), järgmine keskväärtuse 4. ja 5. omadust kasutav võrduste jada:

$$\mathrm{E}(\bar{x}) = \mathrm{E}(\tfrac{1}{n}\sum_{i=1}^{n} x_i) = \tfrac{1}{n}\sum_{i=1}^{n} \underbrace{\mathrm{E}(x_i)}_{\mathrm{E}(X)} = \tfrac{1}{n} n\,\mathrm{E}(X) = \mathrm{E}(X).$$

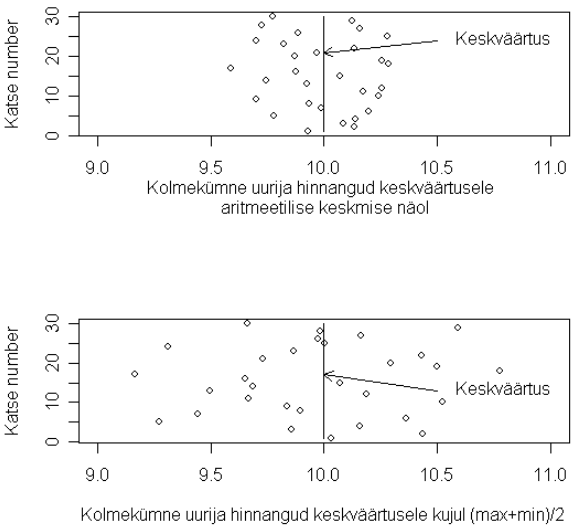
Valimi maksimum saab olla vaid väiksem või võrdne populatsiooni maksimumiga (valimis ei saa olla suuremaid väärtuseid kui populatsioonis, samas võib kuitahes suure valimi korral ikka juhtuda, et populatsioonis „jookseb ringi mõni veel suurema väärtusega indiviid“). Seega on valimi maksimum populatsiooni maksimumi nihkega hinnanguks (vt ka joonis 2.2).

Def. Parameetri θ hinnangut $\hat{\theta}$ nimetatakse **efektiivseks hinnanguks**, kui $\mathrm{var}(\hat{\theta})$ on vähim kõigi parameetri θ nihketa hinnangute dispersioonide hulgas; ehk – efektiivne hinnang on täpseim hinnang.

Näiteks on valimi keskmine populatsiooni keskväärtusele täpsem hinnang võrreldes valimi miinimumi ja maksimumi poolsummaga. Viimane annab ka keskmiselt õige hinnangu populatsiooni keskväärtusele, olles seega nihketa hinnanguks, aga hinnangu varieeruvus on märksa suurem võrreldes aritmeetilise keskmise varieeruvusega (joonis 2.3). See on ka põhjus, miks populatsiooni keskväärtust hinnatakse ikka valmi keskmise ja mitte miinimumi ja maksimumi poolsumma kujul.



Joonis 2.2. Näide nihketa ja nihkega hinnangust.



Joonis 2.3. Näide täpsemast (efektiivsest) ja vähemtäpsemast hindamismeetodist

2.4.4 Hinnangu standardviga

Et andmete alusel leitud parameetri θ hinnang $\hat{\theta}$ on juhuslik suurus, siis eksisteerib tal ka dispersioon $\mathrm{var}(\hat{\theta})$. Viimane on aga jällegi tundmatu üldkogumi parameeter. Seega, et saada tegelikkuses aimu oma andmete alusel leitud parameetri hinnangu täpsusest, tuleb andmetest hinnata ka hinnangu dispersioon, millest reeglina parema mõistetavuse huvides võetakse veel ruutjuur (et saada varieeruvuse hinnangut samal skaalal parameetri endaga).

Def. Hinnangu standardhälbe hinnangut nimetatakse hinnangu **standardveaks**:

$$se(\hat{\theta}) = \sqrt{\hat{\mathrm{var}}(\hat{\theta})}\ . \tag{2.15}$$

Näide. Keskväärtuse E(X) = μ hinnangu $\hat{\mu} = \bar{x}$ dispersiooni hinnang on $\hat{\mathrm{var}}(\hat{\mu}) = s^2/n$ ja standardviga on

$$se(\bar{x}) = \frac{s}{\sqrt{n}},$$

kus s^2 on valimi dispersioon ja s valimi standardhälve.

Saadud valem tuleneb dispersiooni 3.-5. omadusest kujul:

$$\mathrm{var}(\bar{x}) = \mathrm{var}(1/n \sum_{i=1}^{n} x_i) = 1/n^2 \sum_{i=1}^{n} \underbrace{\mathrm{var}(x_i)}_{\mathrm{var}(X)} = 1/n^2\, n\, \mathrm{var}(X) = \frac{\mathrm{var}(X)}{n},$$

millest ruutjuure võtmise ja populatsiooni standardhälbe σ tema hinnanguga (valimi standardhälbega) s asendamise järel on tulemuseks keskmise standardvea valem.

2.5 PRAKTIKUM

2.5.1 Ülesanded

1. Holsteini tõugu veistel domineerib musta värvust määrav alleel (*B*) punast värvust määrava alleeli (*b*) üle ning viimase sagedus populatsioonis on 1/20.

   a) Leidke, kui suur on tõenäosus, et seemendades mustakirjut lehma mustakirju pulli spermaga on tulemuseks punasekirju vasikas (eeldades seejuures, et vanemate eellaste värvuse kohta info puudub)?

   b) Teades, et mustakirju lehma ja mustakirju pulli esimene järglane oli punasekirju lehmvasikas, kui suur on tõenäosus, et sama vanempaari järgmise järglasena sünniks punasekirju pullvasikas? Aga mustakirju lehmvasikas?

2. Piimaveiste suhteline piimajõudluse aretusväärtus SPAV väljendatakse punktides keskväärtusega 100 ja standardhälbega 12 punkti, seejuures eeldatavalt SPAV ~ $N(100, 12)$.

   Leidke,

   a) kui suur peab olema veise SPAV selleks, et ta kuuluks oma tõu 10% paremate loomade hulka;

   b) protsentuaalselt kui suure hulga loomade SPAV on väiksem kui 100 punkti;

   c) protsentuaalselt kui suure hulga loomade SPAV on üle 130 punkti.

   PS. Standardse normaaljaotuse jaotusfunktsiooni väärtus kohal x, $\Phi(x)$, on *MS Excelis* leitav funktsiooniga NORM.S.DIST(x;TRUE) (*Excel* 2003-s funktsiooniga NORMSDIST(x)) ja jaotusfunktsiooni pöördfunktsiooni väärtus kohal p (so arv, millest väiksemaid väärtusi esineb standardse normaaljaotuse puhul tõenäosusega p), $\Phi^{-1}(p)$, funktsiooniga NORM.S.INV(p) (*Excel* 2003-s funktsiooniga NORMSINV(p)).

2.5.2 Ülesannete lahendused

1. a)

   Seemendades mustakirjut lehma mustakirju pulli spermaga, on võimaliku neli alternatiivset ema ja isa genotüüpide kombinatsiooni:

   - esiteks võivad nii ema kui ka isa olla homosügootsed musta värvust määrava alleeli *B* suhtes, so genotüübiga *BB*;
   - teiseks võib ema olla punase geeni kandja heterosügootse genotüübiga *Bb* (või *bB*) ja isa homosügootne genotüübiga *BB*;
   - kolmandaks võib ema olla homosügootne genotüübiga *BB* ja isa punase geeni kandja genotüübiga *Bb* (või *bB*);
   - neljandaks võivad nii ema kui ka isa olla punase geeni kandjad genotüübiga *Bb* (või *bB*).

   Et mõlemad vanemad olid mustakirjud, ei saa nende genotüüp olla *bb* (sellised loomad on punasekirjud).

   Kuna heterosügootsetel genotüüpidel *Bb* ja *bB* vahet ei ole, võib neid käsitleda ühiselt, kasutades näiteks tähistust *Bb*.

   Võttes arvesse mustakirjude vanemate genotüüpide kõikvõimalikud variandid, on järglase punasekirjuna, so genotüübiga *bb*, sündimise tõenäosus avaldatav täistõenäosuse valemi (2.8) abil kujul

   $$\mathrm{P}(bb) = \mathrm{P}(bb \mid BB \times BB)\,\mathrm{P}(BB \times BB) + \mathrm{P}(bb \mid Bb \times BB)\,\mathrm{P}(Bb \times BB) + \mathrm{P}(bb \mid BB \times Bb)\,\mathrm{P}(BB \times Bb) + \mathrm{P}(bb \mid Bb \times Bb)\,\mathrm{P}(Bb \times Bb).$$

   Võrduse paremal pool olevaist liidetavaist on nullist erinev vaid viimane, sest esimese kolme võimaliku vanempaari korral ei saa järglase genotüüp olla *bb* (ehk vastavad tinglikud tõenäosused võrduvad nulliga). Seega $\mathrm{P}(bb) = \mathrm{P}(bb \mid Bb \times Bb)\,\mathrm{P}(Bb \times Bb)$.

Kui punast värvust määrava alleeli *b* sagedus populatsioonis on 1/20, siis on musta värvust määrava alleeli *B* sagedus P(*B*) = 1 – P(*b*) = 19/20 ja heterosügootsete loomade (genotüübiga *Bb* punase geeni kandjate) osakaal populatsioonis on Hardy-Weinbergi seaduse eeldusel P(*Bb*) = 2×(19/20)×(1/20) = 19/200.

Eeldades, et vanemate valikul ei arvestata nende genotüüpi, on ema ja isa heterosügootseks osutumise näol tegu sõltumatute sündmustega, mistap avaldub heterosügootse vanempaari saamise tõenäosus kujul

P(*Bb* × *Bb*) = P(*Bb*) × P(*Bb*) = 19/200 × 19/200 = 361/40000 = 0,009025.

Vastavalt Mendeli II seadusele P(*bb* | *Bb* × *Bb*) = ¼ = 0,25.

Kokkuvõttes:

P(*bb*) = P(*bb* | *Bb* × *Bb*) P(*Bb* × *Bb*) = 0,009025 × 0,25 ≈ 0,00226.

Seega peaks 1000-st mustakirjude vanemate järglasest keskmiselt 2-3 olema punasekirjud.

b)

Teadmine, et mustakirju lehma ja mustakirju pulli esimene järglane oli punasekirju vasikas, ütleb, et mõlemad vanemad pidid olema punast värvust määrava alleeli kandjad genotüübiga *Bb*.

Tõenäosus, et sama vanempaari järgmise järglasena sünnib punasekirju, so genotüübiga *bb*, vasikas, avaldub vastavalt Medeli II seadusele kujul P(*bb* | *Bb* × *Bb*) = ¼.

Tõenäosus, et sünnib pullvasikas, on ½.

Et vasika värvus ja sugu on sõltumatud, samuti ei sõltu vasika sugu vanemate genotüübist, on punasekirju pullvasika sündimise tõenäosus

P(*bb* ∩ pullvasikas | *Bb* × *Bb*)
= P(*bb* | *Bb* × *Bb*) P(pullvasikas | *Bb* × *Bb*) = P(*bb* | *Bb* × *Bb*) P(pullvasikas) = ¼ × ½ = ⅛.

Mustakirju lehmvasikas (genotüübiga *BB* või *Bb*) sünnib vastavalt Mendeli seadustele ning soo ja vanemate genotüübi sõltumatuse eeldusel tõenäosusega

P[(*BB* ∪ *Bb*) ∩ lehmvasikas | *Bb* × *Bb*]
= [P(*BB* | *Bb* × *Bb*) + P(*Bb* | *Bb* × *Bb*)] P(lehmvasikas) = (¼ + ½) × ½ = ¾ × ½ = ⅜.

2. a) Otsitav suurus *x* on matemaatiliselt avaldatav seosena P(SPAV > *x*) = 0,1.

Vastavalt jaotusfunktsiooni omadustele kehtib siis ka seos P(SPAV < *x*) = *F*(*x*) = 0,9.

Kuna SPAV ~ *N*(100, 12), siis (SPAV–100)/12 ~ *N*(0, 1) ja *F*(*x*) = Φ[(*x*–100)/12] = 0,9.

Standardse normaaljaotuse jaotusfunktsiooni argument *z* = (*x*–100)/12, mis vastab jaotusfunktsiooni väärtusele Φ(*z*) = 0,9, on 1,28 (leitav on see näiteks *Excelis* funktiooniga =NORM.S.INV(0,9)).

Seega *z* = (*x*–100)/12 = 1,28, millest *x* = 100 + 1,28×12 = 115,4.

Vastus: oma tõu 10% paremate loomade hulka kuulumiseks peab SPAV olema üle 115 punkti.

b) P(SPAV < 100) = ?

Kuna SPAV ~ *N*(100, 12) ja normaaljaotuse puhul langevad keskmine ja mediaan kokku, on 50% loomade SPAV väiksem kui 100 punkti ja 50% loomade SPAV suurem kui 100 punkti. St, et P(SPAV < 100) = 0,5.

c) P(SPAV > 130) = ?

P(SPAV > 130) = 1 – P(SPAV < 130) = 1 – P[(SPAV–100)/12 < (130–100)/12] = 1 – Φ(2,5)
= 1 – 0,994 = 0,006.

Seega peaks enam kui 130-punktlise SPAV-ga olema eeldatavalt 0,6% Eesti piimaveiseid.