2.3 Meetodeid pidevate tunnuste analüüsiks

Kui laps sarnaneb isaga, on tegemist pärilikkusega.
Kui ta sarnaneb naabrimehega,
on tegemist väliskeskkonna mõjudega.

Murphy seadus pärilikkuse kohta

Pidevate arvtunnuste (kaal, vererõhk, eluiga peale haiguse tuvastamist jne meditsiinis ning suurem osa jõudlust iseloomustavaid tunnuseid loomakasvatuses) korral pole enamasti võimalik indiviidide vahelisi erinevusi kirjeldada vaid ühe konkreetse geeni mõjuga. Vastuolu tekib sellest, et genotüüp ühes konkreetses lookuses on oma olemuselt diskreetne ja ei saa seega üksinda põhjustada indiviididevahelisi erinevusi pideva tunnuse osas. See, et pidevate arvtunnuste jaotuseks populatsioonis on enamasti normaaljaotus, vihjab nende tunnuste variatsiooni tingitusele paljude geenide efektidest ja ka keskkonnamõjudest.

Seetõttu püütaksegi pidevate arvtunnuste geneetilisel uurimisel esmalt selgeks teha, kuivõrd ja kas üldse on selle tunnuse väärtused geenide poolt määratud ja kas viimaste mõju ka järglastele edasi kandub. Lihtsaim viis seda selgitada on võrrelda omavahel tunnuse väärtusi (näiteks keskmist vererõhku) erinevates "perekondades", püüdes seejuures arvestada ka võimalikku mittegeneetilist nn keskkonna mõju.

2.3.1 Kaksikutemeetod

Lihtsaim meetod selgitamaks uuritava tunnuse geneetilist determineeritust on inimgeneetikas aastakümneid rakendatud F. Caltoni poolt 19. sajandi lõpul välja töötatud **kaksikutemeetod**, mis baseerub faktil, et ühemunarakukaksikud on geneetiliselt identsed, kahemunarakukaksikuil on aga sarnaselt täisõvedega ühised vaid keskmiselt pooled geenidest.

Eeldades, et kaksikud kasvavad üles samades tingimustes (st et keskkonnamõjud uuritavale tunnusele on võrdsed), peaks ühemunarakukaksikute suurem sarnasus võrreldes kahemunarakukaksikutega olema tingitud vaid nende suuremast ühiste geenide hulgast.

Seega on ühemunarakukaksikute suurem sarnasus võrreldes kahemunarakukaksikutega tõend uuritava tunnuse geneetilisest määratusest.

2.3.2 Biomeetriline geneetika

Looma- ja taimekasvatuses on majandusliku kasu huvides püütud alati valida järglaspõlvkonna vanemaiks geneetiliselt potentsiaalilt parimaid isendeid. Nii nende isendite kui ka nende poolt järglastele pärandatava efekti välja selgitamiseks kasutatakse vägagi erinevaid mudeleid, kus geneetilisele efektile püütakse enamasti jälile saada erinevate sugulaste gruppide omavahelise võrdlemise teel dispersioonanalüüsi abil. Et huvi pakub just see, kui suur osa uuritava tunnuse varieeruvusest on tingitud erinevate indiviidide erinevast genotüübist, siis loetakse geneetiline efekt (geneetilise varieeruvuse allikas – isa, ema, indiviid ise jne) enamasti juhuslikuks ja analüüs baseerub dispersioonanalüüsi segamudelite teoorial.

¤ Päritavus, geneetiline mudel

Geneetiliste struktuuride (kromosoomide, geenide) ja neis sisalduva geneetilise informatsiooni ülekannet vanematelt järglastele (või emarakult tütarrakkudele), mille tulemusena kujunevad viimaste genotüübid, nimetatakse **pärandumiseks**.

Seda, kui suur osa mingi populatsiooni fenotüübilisest muutlikkusest on tingitud sinna kuuluvate isendite geneetilisest (pärilikust) erinevusest, kirjeldab **päritavuskoefitsient** (tähist. h^2).

Päritavuskoefitsiendi hindamiseks esitatakse indiviidi fenotüübiväärtus P tema genotüübiväärtuse G ja juhuslike keskkonnamõjude E summana:

$$P = G + E\,.$$

Fenotüübiline varieeruvus on siis populatsiooni geneetilise varieeruvuse ja keskkonnatingimustest tingitud varieeruvuse summa:

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2\,,$$

ning päritavuskoefitsient geneetilise dispersiooni suhe tunnuse kogudispersiooni:

$$h^2 = \sigma_G^2 / \sigma_P^2 .$$

Päritavuse sisu õigeks mõistmiseks on vaja arvestada, et päritavuskoefitsient hindab tunnuse geneetilise muutlikkuse osa antud geneetilise struktuuriga populatsioonis konkreetsetes keskkonnatingimustes ega näita tunnuse päriliku tingituse määra ja mehhanismi üksikindiviidide arengus.

Arvestades, et genotüübiväärtus *G* sisaldab eneses nii üksikute geenide kui ka nende kõikvõimalike koosmõjude efekte, millest viimased järglastele edasi ei kandu, esitatakse päritavuskoefitsient sageli ka vaid üksikute geenide aditiivsest mõjust tingitud varieeruvuse σ_A^2 ja koguvarieeruvuse σ_P^2 suhtena (nö päritavus kitsamas mõistes):

$$h^2 = \sigma_A^2 / \sigma_P^2 .$$

Indiviidi üksikute geenide aditiivset mõju, mida tähistatakse tähega *A* ja millest pool pärandub järglastele, nimetatakse selle indiviidi **aretusväärtuseks**.

Päritavuskoefitsient saab omada väärtusi 0-st 1-ni (100%-ni). Mida suurem on päritavuskoefitsiendi väärtus, seda sarnasemad on ühte geneetilisse gruppi kuuluvad indiviidid. Viimane vihjab aga otseselt tunnuse (haiguse) tugevale sõltuvusele pärilikust geneetilisest materjalist.

¤ **Ülevaade geneetiliste parameetrite hindamisel kasutatavatest mudelitest**

Päritavuskoefitsiendi hindamine tänapäeval baseerub dispersioonanalüüsi segamudelitel, kus võimalikult täpselt püütakse fikseerida nii uuritavate isendite omavaheline sugulus kui ka uuritavat tunnust mõjutavad keskkonnafaktorid. Et pidevate arvtunnuste geneetilise determineerituse selgitamine on aastakümneid olnud oluline just loomakasvatuses seisukohast, on enamus mudeleid välja töötatud põllumajanduse intensiivistamise ja suurema kasu – mille tagab aga just geneetiliste parameetrite täpsem hindamine –, huvides.

Antud valdkonna põhjalikum käsitlemine ei mahu käesoleva kursuse raamidesse. Põhjalikuma teabe saamiseks vt näiteks:

Kaart, T. (2001) Ülevaade geneetiliste parameetrite hindamisel kasutatavatest mudelitest. Eesti Põllumajandusülikooli Loomakasvatusinstituudi teadustöid, 71: lk 52-66, Tartu, LKI (http://www.eau.ee/~ktanel/kaart_2001_LKI.pdf).

Kaart, T. (2008). EMÜ VLI loomakasvatuse ja -geneetika eriala magistrantidele mõeldud suhteliselt matemaatiline loengukursus „Loomade aretusväärtuse hindamine ja aretusprogrammid“, http://www.eau.ee/~ktanel/VL_0192/

Schaeffer, L. R. Guelphi Ülikoolis loetavate loengukursuste „*Quantitative Genetics and Animal Models*“ ning „*Estimation of Genetic Parameters*“ põhjalikud loengukonspektid, http://www.aps.uoguelph.ca/~lrs/Animals/

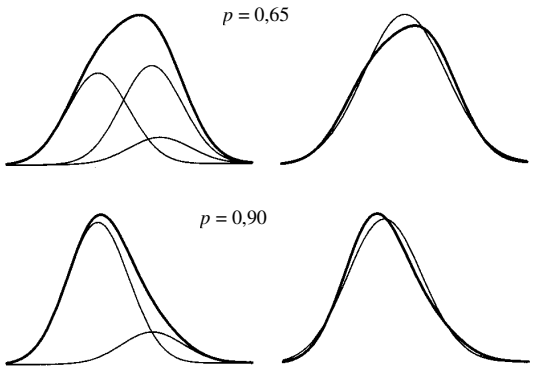
2.3.3 Elementaarsed testid üksikute suure mõjuga geenide tuvastamiseks

Olles jõudnud selgusele tunnuse suuremal või vähemal määral geneetilises determineerituses kerkib kohe ka uus probleem – kui suure hulga geenide poolt selle tunnuse potentsiaalne avaldumine ära määratakse? Mingi haiguse ravi oleks ju tunduvalt lihtsam välja töötada, kui haigestumine sõltuks vaid ühest konkreetsest DNA piirkonnast (geenist) ja selle järgi produtseeritavatest (või produtseerimata jäetavatest) valkudest. Ka põllumajanduse seisukohast on kasulik, kui mingi majanduslikult oluline tunnus vaid väheste geenide poolt ära määratakse – näiteks loomakasvatuses muudaks vaid nö heade geenidega vanemloomade valik aretuse tunduvalt efektiivsemaks ja taimekasvatuses on juba saavutatud tõeline murrang, siirdades varem külma- või haiguskartlike sortide genoomi külmakindlust või haigusresidentsust suurendavaid DNA osi.

¤ **Tunnuse jaotuse erinevus normaaljaotusest**

Selles, kas huvipakkuv tunnus on seotud vaid üksikute geenidega, on aastakümneid püütud selgusele jõuda, võrreldes tunnuse empiirilist jaotust normaaljaotusega. On ju normaaljaotus paljude väikese mõjuga geenide summaarsel toimel kujunevate tunnuste mudeljaotuseks ja erinevus sellest mudelist vihjab mingite eelduste mittetäidetusele – näiteks mõne üksiku geeni prevaleerimisele teiste üle.

Kontrollimaks uuritava tunnuse jaotumist vastavalt normaaljaotusele on kasutatavad kõik rohkem või vähem traditsioonilised selleks otstarbeks konstrueeritud testid. Lihtsaimad ja ka üldtuntud viisid on näiteks graafiline võrdlemine (tõenäosuspaber) ja Kolmogrov-Smirnovi test. Tavaliselt ei ole aga analüüsitavad andmed puhtad kõikvõimalikest kõrvalmõjudest ja kui nad ka on, ei pruugi erinevate genotüüpidega isendite summaarne jaotus normaalsest palju erineda (vt näiteks joonist 11).



Joonis 11. Üksiku dialleelse lookuse poolt määratud tunnuse fenotüübijaotused, kus igale genotüübile vastav fenotüüp jaotub normaaljaotuse järgi sama dispersiooniga σ^2. Tunnuse keskmised väärtused erinevate genotüüpide korral on $\mu_{qq} = -\sigma$, $\mu_{Qq} = 0{,}75\sigma$ ja $\mu_{QQ} = \sigma$ ning genotüübid on Hardy-Weinbergi tasakaalus alleeli q sagedusega p = 0,65 (ülal) ja p = 0,9 (all). Vasakpoolsetel joonistel on kujutatud kolmele genotüübile vastavad jaotused ja nende segujaotus, parempoolsetel on võrreldud saadud segujaotust samade parameetritega normaaljaotusega.

¤ Õvedevahelisel erinevusel baseeruvad testid

Nende testide korral eeldatakse, et ühe suure mõjuga geeni korral kasvab alleelide lahknemise tagajärjel perekonnasisene fenotüübiline varieeruvus. Näiteks kui alleelid Q ja q märgivad kahte suure mõjuga geeni esinemise vormi, siis peaks varieeruvus heterosügootsete vanemate ($Qq \times Qq$) järglaste hulgas olema suurem, kui homosügootide ($QQ \times QQ$ ja $qq \times qq$) korral; genotüüpide $QQ \times Qq$ ja $qq \times Qq$ ristamise tulemusel saadud järglaste vaheline varieeruvus peaks aga jääma kahe eelnevalt kirjeldatud grupi vahele. Kontrolliks sobivad gruppidevahelise varieeruvuse erinevuse tõestamiseks mõeldud testid (normaaljaotuse korral näiteks Bartlett' test).

Teine lähenemine sarnastele andmetele põhineb mõttekäigul, et kui tunnus on määratud paljude väikese ja võrdse mõjuga geenide poolt, siis ei tohiks ühe perekonna järglaste vaheline erinevus sõltuda nende vanemate fenotüübist, ja kui uuritav tunnus on siiski määratud väheste suure mõjuga geenide poolt, peaks teistest palju erineva (ekstreemse) fenotüübiga vanemad olema ilmselt homosügoodid ja keskmise fenotüübiga isendid heterosügoodid. See mõttekäik viib ühe perekonna järglaste sisese varieeruvuse ja nende vanemate keskmise fenotüübiväärtuse vahelise regressiooniseoseni:

$$\mathrm{D}(z_i) = a + b_1\overline{z}_i + b_2\overline{z}_i^2,$$

kus $\mathrm{D}(z_i)$ on i. vanempaari (perekonna) järglaste vaheline dispersioon ja $\overline{z}_i$ on i. vanempaari keskmine fenotüübiväärtus (viimase mitteteadmisel võib selle asendada ka vastavate järglaste keskmise fenotüübiväärtusega). Üksiku märkimisväärse mõjuga geeni olemasolule vihjab ruutliikme ees oleva regressioonikordaja b_2 statistiliselt oluliseks osutumine (ruutliiget vaadatakse põhjusel, et lihtne lineaarne seos võib olla tingitud paljudele tunnustele omasest suuremate fenotüübiväärtustega kaasnevast suuremast varieeruvusest).

¤ Mitteparameetrilised liinide ristamise testid

Oletame, et uuritava tunnuse potentsiaalne väärtus on täielikult ära määratud ühe lookuse kahe alleeli, Q ja q, poolt. Iga genotüübiga isendeile mõjub ka keskkond, mille efektide jaotus võib genotüübiti erineda. Kogu populatsiooni fenotüübijaotuse võib esitada kolmele genotüübile vastavate fenotüübijaotuste kaalutud seguna:

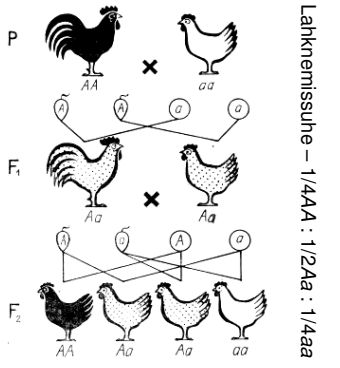
$$p(z) = p_{QQ}(z)\,\mathrm{P}(QQ) + p_{Qq}(z)\,\mathrm{P}(Qq) + p_{qq}(z)\,\mathrm{P}(qq), \qquad (2.4)$$

kus P(QQ) on tõenäosus, et juhuslikult valitud indiviid on genotüübiga QQ ja $p_{QQ}(z)$ on genotüübiga QQ isendile mõjuvate keskkonnaefektide jaotus (fenotüübijaotus); ülejäänud muutujad on defineeritud analoogselt.

Järgnevalt vaadeldakse kahte vanempopulatsiooni, P_1 ja P_2, millest üks on homosügootne alleeli Q ja teine alleeli q suhtes. Kui nüüd järglaste erinevus põlvkonnas F_2 on tingitud vaid konkreetses lookuses paiknevate alleelide erinevusest, avaldub F_2-populatsiooni jaotus valemi (2.4) järgi, kus $p_{Qq}(z) = p_{F_1}(z)$, $p_{QQ}(z) = p_{P_1}(z)$, $p_{qq}(z) = p_{P_2}(z)$, P(Qq) = ½ ja P(QQ) = P(qq) = ¼ (vt ka joonist 12):

$$p_{F_2}(z) = \frac{p_{P_1}(z)}{4} + \frac{p_{F_1}(z)}{2} + \frac{p_{P_2}(z)}{4}.$$

Hüpoteesina kontrollitakse siis uuritava tunnuse empiirilise jaotuse samasust tuletatud teoreetilise jaotusega.



Joonis 12. Homosügootsete vanemate kaks järglaspõlvkonda.

2.4 Kompleksne segregatsioonanalüüs

2.4.1 Segregatsioonanalüüsi segumudelid

Praktiliste probleemide lahendamisel osutub eelnevalt kirjeldatud testide võimsus enamasti liialt väikeseks (liialt palju informatsiooni jääb kasutamata). Seetõttu rakendatakse ka pidevate arvtunnuste korral päritavuse seaduspärade selgitamisel ja põhigeenide otsimisel fenotüübi- ning põlvnemis-andmete alusel märksa võimsamat segregatsioon- ehk lahknemisanalüüsi. Selle, nn **komplekssе segregatsioonanalüüsi** korral, püütakse uuritava tunnuse fenotüübijaotust lähendada erinevatele genotüüpidele vastavate jaotuste seguga, kus segu moodustavate genotüübijaotuste proportsioonid hinnatakse põlvnemis- ja fenotüübiandmetele tuginedes. Enamasti, nagu statistilises andmeanalüüsis ikka, eeldatakse nii fenotüübi- kui ka genotüübijaotuste vastamist normaalsuse eeldustele.

¤ **Jaotuste segu**

Kujutagu uuritava tunnuse fenotüübijaotus enesest kaalutud genotüübijaotuste summat. Kui meil on $i = 1,...,m$ genotüübijaotust, $p_1(z),..., p_m(z)$, igaüks sagedusega P(i), siis tunnuse z tihedusfunktsioon $p(z)$ avaldub kujul

$$p(z) = \sum_{i=1}^{m} \mathrm{P}(i) p_i(z) .$$

Eeldades genotüüpide jaotumist vastavalt normaaljaotusele saame, et

$$p(z) = \sum_{i=1}^{m} \mathrm{P}(i) f(z, \mu_i, \sigma_i^2) , \qquad (2.5)$$

kus

$$f(z, \mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(z-\mu_i)^2}{2\sigma_i^2} \right].$$

Fenotüübijaotus (2.5) sisaldab kokku $3m-1$ hindamist vajavat parameetrit: $m-1$ kaaluparameetrit P(i) ning m genotüübijaotuste keskväärtust μ_i ja dispersiooni σ_i^2. Kuna tavaliselt eeldatakse genotüüpide võrdset varieeruvust, st $\sigma_1^2 = ... = \sigma_m^2 = \sigma^2$, kahaneb hinnatavate parameetrite arv $2m$-ni.

¤ **Parameetrite hindamine**

Jaotuse (2.5) parameetrite hindamiseks kasutatakse enamasti suurima tõepära (ML – *maximum likelihood*) meetodit, mille korral teoreetilise jaotuse (2.5) sõltuvus tundmatutest parameetritest P(1),…,P(m), $\mu_1,\ldots,\mu_m$, $\sigma_1^2,\ldots,\sigma_m^2$ esitatakse **tõepärafunktsioonina** $L(z)$, kus argumendi rollis on vaadeldud fenotüübiväärtused.

Näitena tõepärafunktsiooni konstrueerimisest vaatleme situatsiooni, kus analüüsitavad indiviidid on juhuslikult valitud populatsioonist, milles uuritava tunnuse potentsiaalne väärtus on määratud ühe dialleelse lookusega. Võttes genotüüpideks $i = QQ, Qq, qq$ ning eeldades genotüübijaotuste normaal-sust parameetritega μ_i ja σ^2, saame j. indiviidi jaoks välja kirjutada tõepärafunktsiooni $L(z_j)$ (vrd valem (2.4)):

$$L(z_j) = \mathrm{P}(QQ) f(z_j, \mu_{QQ}, \sigma^2) + \mathrm{P}(Qq) f(z_j, \mu_{Qq}, \sigma^2) + \mathrm{P}(qq) f(z_j, \mu_{qq}, \sigma^2) .$$

Üldine tõepärafunktsioon on suguluses mitteolevate indiviidide korral esitatav üksikute indiviidide jaoks välja kirjutatud tõepärafunktsioonide korrutisena:

$$L(\mathbf{z}) = L(z_1, z_2,..., z_n) = \prod_{j=1}^{n} L(z_j) ,$$

n – vaadeldud indiviidide arv.

Analoogselt on tõepärafunktsioon esitatav ka keerulisemate geneetiliste mudelite korral – iga üksiku indiviidi jaoks välja kirjutatud tõepärafunktsiooni avaldis koosneb siis lihtsalt enamaist liidetavatest (iga võimalik genotüüp annab ühe liidetava).

Eeldades Hardy-Weinbergi seaduse kehtimist vaatluse all olevas populatsioonis, väheneb hindamist vajavate parameetrite arv meie näites viieni: alleeli q (või Q) sagedus p (geneetilise tasakaalu kehtides on sellest avaldatavad kõik vajalikud sagedused P(i), $i = QQ, Qq, qq$), genotüübijaotuste dispersiooni-parameeter σ^2 ja keskväärtused μ_{QQ}, μ_{Qq} ja μ_{qq}.

Parameetrite suurima tõepära hinnangud on parameetrite väärtused, mille korral fikseeritud fenotüübi-andmetega $\mathbf{z} = (z_1, z_2, ..., z_n)$ tõepärafunktsiooni väärtus on maksimaalne. Hinnangute saamiseks on välja töötatud rida numbrilisi ja iteratiivseid meetodeid. Üks enamkasutatavaid hindamismeetodeid on EM-algoritm (*expextation-maximization method*), selle rakendamise konkreetse näiteni segregatsioon-analüüsil jõuame mõne punkti pärast.

¤ Hüpoteeside kontrollimine

Üks olulisemaid küsimusi segregatsioonanalüüsil on mudeli valik – milline on optimaalne liidetavate (genotüüpide) arv uuritava (fenotüübil kirjeldatud) tunnuse lähendamiseks ja kuidas erinevaid mudeleid (jaotuste segusid) võrrelda. Näiteks võime oma andmetele esmalt sobitada kahe normaaljaotuse järgi jaotuva liidetavaga ning seega neljast parameetrist (μ_1, μ_2, σ^2 ja p) sõltuvat segumudelit. Kas aga lähendab selline neljast parameetrist sõltuv jaotuste segu meie andmeid oluliselt paremini, kui lihtsalt kahe parameetriga, μ ja σ^2, määratud normaaljaotus? Lihtsaim teststatistik selle kontrollimiseks on **tõepärasuhte statistik**

$$\Lambda(\mathbf{z}) = -2\ln\left[\frac{\hat{L}_k(\mathbf{z})}{\hat{L}(\mathbf{z})}\right] = -2\left\{\ln\left[\hat{L}_k(\mathbf{z})\right] - \ln\left[\hat{L}(\mathbf{z})\right]\right\},$$

kus $\hat{L}(\mathbf{z})$ on keerulisemale mudelile vastava tõepärafunktsiooni maksimumi hinnang ja $\hat{L}_k(\mathbf{z})$ on kitsendatud mudelile (r keerulisema mudeli parameetrit loetakse fikseerituiks) vastava tõepärafunktsiooni maksimumi hinnang. Piisavalt suure valimi korral on tõepärasuhte statistik χ^2-jaotusega, vabadusastmete arvuga r.

Lihtsaim kitsendatud mudel eeldab, et tihedusfunktsiooniga (2.5) määratud jaotus kujutab enesest tavalist normaaljaotust tundmatute parameetritega μ ja σ^2 (st et uuritava tunnuse potentsiaalne väärtus on määratud paljude geenide väikeste ja võrdsete mõjude poolt – puudub teistest oluliselt suurema mõjuga geen). Statistika teooriast on teada, et n sõltumatu normaaljaotusega $N(\mu,\sigma^2)$ vaatluse korral kujutab vaatlusvektori tõepärafunktsioon enesest n-i identse, parameetritest μ ja σ^2 sõltuva normaaljaotuse tihedusfunktsiooni korrutist. Parameetrite suurima tõepära hinnangud on

$$\hat{\mu} = \bar{z} \text{ ja } \hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(z_j - \bar{z})^2$$

ning neile vastav tõepärafunktsiooni maksimum

$$\hat{L}_r(\mu, \sigma^2, \mathbf{z}) = \prod_{j=1}^{n}\left\{\frac{1}{\sqrt{2\pi\hat{\sigma}^2}}\exp\left[-\frac{(z_j - \hat{\mu})^2}{2\hat{\sigma}^2}\right]\right\}.$$

Võttes viimasest avaldisest logaritmi ja kirjutades lahti parameetrite hinnangud, saame

$$-2\ln\left[\hat{L}_r(\mu, \sigma^2, \mathbf{z})\right] = n\left[\ln(\hat{\sigma}^2) + \ln(2\pi) + 1\right]. \tag{2.6}$$

Keerulisema mudeli korral pole tõepärafunktsiooni maksimumi avaldamine enam nii lihtne. Näitena vaatleme jällegi eeldatavalt ühe lookuse kahe alleeli poolt määratud tunnust.

Näide. Kasutame tõepärasuhte statistikut kontrollimaks, kas dialleelse põhigeeni eeldusel konstrueeritud segujaotus lähendab andmeid paremini kui tavaline normaaljaotus. Segujaotuse puhul eeldame, et uuritav populatsioon on Hardy-Weinbergi tasakaalus ja et oletatava põhigeeni baasil kujunevad 3 genotüübijaotust on konstantse dispersiooniga normaaljaotused. Koosnegu valim n-st omavahel mitte suguluses olevast juhuslikult valitud indiviidist. Tõepärasuhte statistik esitub siis kujul

$$\Lambda(\mathbf{z}) = -2\left\{\ln\left[\hat{L}_k(\mathbf{z})\right] - \ln\left[\hat{L}(\mathbf{z})\right]\right\} = 2\ln\left[\hat{L}(\mathbf{z})\right] + n\left[\ln(\hat{\sigma}^2) + \ln(2\pi) + 1\right],$$

kus

$$\hat{L}(\mathbf{z}) = \max\left[\prod_{j=1}^{n}\hat{L}(z_j)\right].$$

Maksimeerimine toimub üle kõigi lubatavate parameetrite väärtuste, $0 \le p \le 1$, $-\infty < \mu_{QQ}, \mu_{Qq}, \mu_{qq} < \infty$, $\sigma^2 \ge 0$ ning

$$L(z_j) = p^2 \times f(z_j, \mu_{QQ}, \sigma^2) + 2p(1-p) \times f(z_j, \mu_{Qq}, \sigma^2) + (1-p)^2 \times f(z_j, \mu_{qq}, \sigma^2).$$

Kuna täielik mudel sisaldab viit parameetrit, kitsendatud mudel aga kahte, on teststatistik

$\Lambda(\mathbf{z})$ ligikaudu χ^2-jaotusega, vabadusastmete arvuga 3. Kolme normaaljaotuse segujaotus lähendab andmeid paremini kui tavaline normaaljaotus juhul, kui tõepärasuhte statistiku väärtus on suurem kui 7,82 või 11,4 (vastavalt olulisuse nivoo 0,05 või 0,01 korral).

`

Tõepärasuhte statistiku puuduseks on, et see eeldab mudelite allutatust – üks võrreldav mudel saadakse teisest osade parameetrite fikseerimise tulemusel. Kui mudelid ei ole allutatud, ei ole kuitahes suure valimi korral statistiku jaotuseks χ^2-jaotus. Sellisel juhul on tõepärafunktsioonide võrdlemiseks kasutatav **Akaike informatsiooni kriteerium**

$$\text{AIC} = -2\ln(\text{tõepärafunktsiooni väärtus}) + 2(\text{parameetrite arv}).$$

Väiksem AIC vastab sobivamale mudelile. AIC puuduseks on, et ta ei võimalda kontrollida hüpoteese ühe või teise mudeli paremast sobivusest, eeliseks aga võimalus omavahel võrrelda vägagi erinevaid mudeleid.

Peale kahe kirjeldatud meetodi võib mudelite võrdlemisel kasutada ka taasvaliku meetodeid (*resampling methods*) nagu näiteks permutatsioonitestid ja *bootstrap*-meetodid.

2.4.2 Kompleksse segregatsioonanalüüsi mudelid

Eelnevalt kirjeldatud segumudelite (-jaotuste) teooria fikseerib segregatsioonanalüüsi matemaatilised alused. Geneetika ülesandeks on defineerida alternatiivsed omavahel võrreldavad uuritava tunnuse geneetilise determineerituse laadid (mudelid), mis määravad ära segukomponentide arvu ja proportsioonid. Enne mõningate mudelite lähemat tutvustamist vaatleme näite abil, millist tüüpi parameetrite hindamisega ja milliste hüpoteeside kontrollimisega kompleksne segregatsioonanalüüs tegeleb.

Näide. Vaatleme mudelit, mis võtab arvesse nii võimaliku kahealleelse põhigeeni (alleelid *Q* ja *q*) mõju kui ka paljude väikese mõjuga geenide taustaefekti. Fenotüübiväärtused eeldatakse olevat põhigeeni suhtes tingliku normaaljaotusega, keskväärtustega vastavalt μ_{QQ}, μ_{Qq} ja μ_{qq} ning ühise dispersiooniga $\sigma^2 = \sigma_E^2 + \sigma_A^2$ (keskkonnadispersioon pluss taustaefekti moodustavate väikeste geenimõjude aditiivdispersioon). Hardy-Weinbergi seaduse kehtides sisaldab kirjeldatud mudel kuut hinnatavat parameetrit: alleeli *Q* sagedus *p*, genotüüpidele *QQ*, *Qq* ja *qq* vastavad keskväärtused, keskkonnamõjust tingitud varieeruvus σ_E^2 ning väikses mõjuga geenide summaarsest efektist tingitud varieeruvus σ_A^2. Erinevad võrreldavad mudelid koos vabade ja fikseeritud parameetritega on toodud järgnevas tabelis.
Soovides kontrollida hüpoteesi suure mõjuga geeni olemasolu kohta, tuleb omavahel võrrelda mudeleid 1 ja 2, neist esimene on kitsendatud ja teine analüüsi mõttes täielik mudel. Tõepärasuhte statistiku vabadusastmete arv on kolm.

Mudel	Hinnatavad parameetrid	Fikseeritud parameetrid
1. Geneetilised efektid puuduvad	μ, σ_E^2	$\mu_{QQ} = \mu_{Qq} = \mu_{qq} = \mu$, $p = 0$, $\sigma_A^2 = 0$
2. Põhigeen, väikeste geenimõjude taustaefekt puudub	μ_{QQ}, μ_{Qq}, μ_{qq}, p, σ_E^2	$\sigma_A^2 = 0$
3. Väikese mõjuga geenide summaarne efekt, põhigeen puudub	μ, σ_E^2, σ_A^2	$\mu_{QQ} = \mu_{Qq} = \mu_{qq} = \mu$, $p = 0$
4. Täielik mudel: põhigeen pluss geneetiline taustaefekt	μ_{QQ}, μ_{Qq}, μ_{qq}, p, σ_E^2, σ_A^2	puuduvad

¤ **Tõepärafunktsioonid ühe suure mõjuga geeni korral**

Oletame, et uuritava tunnuse väärtuse määrab ära kahe esinemisvormiga põhigeen. Vaatleme *i*. perekonna (vanemate paari) *j*. järglast o_{ij} ning tema isa f_i ja ema m_i (järgnevais valemeis on lühiduse huvides kasutatud tähistusi o_j, *f* ja *m*, mis muidugi ei tähenda, et need suurused vaatlusaluse perekonna muutudes samaks jääksid). Järglase fenotüübiväärtust tähistame z_{ij}. Genotüübid analüüsitava lookuse suhtes on indekseeritud järgmiselt: $g = 1$, kui genotüübiks on *QQ*, $g = 2$, kui genotüübiks on *Qq* ja $g = 3$, kui genotüübiks on *qq*; seejuures märgivad g_f, g_m ja g_{o_j} vastavalt isa, ema ja nende *j*. järglase

genotüüpi. Igale genotüübile vastav fenotüüp eeldatakse jaotuvat normaaljaotuse järgi keskväärtusega μ_g ja dispersiooniga σ^2.

Sõltuvana vanemate genotüüpidest avaldub järglase *ij* tõepärafunktsioon seosena`

$$L(z_{ij} \mid g_f, g_m) = \sum_{g_o=1}^{3} \mathrm{P}(g_o \mid g_f, g_m) \times f(z_{ij}, \mu_{g_o}, \sigma^2) \tag{2.7}$$

kus $\mathrm{P}(g_o \mid g_f, g_m)$ märgib tõenäosust, et järglane omab genotüüpi g_o, kui tema vanematel on vaadeldavas lookuses geenikomplektid g_f ja g_m (so **lahknemistõenäosus**). Need tõenäosused kujutavad enesest tingliku tõepärafunktsiooni (2.7) kujul oleva segumudeli kaaluparameetreid ja on leitavad vastavalt Mendeli lahknemisseadustele.

Näiteks kui isa ja ema genotüübid tunnuse kujunemist kvalitatiivses mõttes determineeriva geeni suhtes on vastavalt *QQ* ja *Qq*, ehk $g_f = 1$ ja $g_m = 2$, siis

$$\begin{aligned} &\mathrm{P}(g_o = 3 \mid g_f = 1, g_m = 2) = \mathrm{P}(qq \mid g_f = QQ, g_m = Qq) = 0, \\ &\mathrm{P}(g_o = 2 \mid g_f = 1, g_m = 2) = \mathrm{P}(Qq \mid g_f = QQ, g_m = Qq) = 1/2, \\ &\mathrm{P}(g_o = 1 \mid g_f = 1, g_m = 2) = \mathrm{P}(QQ \mid g_f = QQ, g_m = Qq) = 1/2 \end{aligned} \tag{2.8}$$

ning valem (2.7) lihtsustub:

$$L(z_{ij} \mid QQ, Qq) = \tfrac{1}{2} f(z_{ij}, \mu_{QQ}, \sigma^2) + \tfrac{1}{2} f(z_{ij}, \mu_{Qq}, \sigma^2) \tag{2.9}$$

Määratuna vaid vanemate genotüüpide poolt, on iga järglane sõltumatu oma õdedest-vendadest ja seega on *i*. perekonna jaoks väljakirjutatud tõepärafunktsioon üksikute järglaste tõepärafunktsioonide (2.7) korrutis:

$$L(z_{i\cdot} \mid g_f, g_m) = \prod_{j=1}^{n_i} L(z_{ij} \mid g_f, g_m) \tag{2.10}$$

n_i – järglaste arv *i*. perekonnas.

Kuna me tegelikult vanemate genotüüpe ei tea, esitatakse perekonnale *i* vastav tõepärafunktsioon kui summa üle vanemate kõigi võimalike genotüüpide paaride (mida antud juhul on üheksa):

$$L(z_{i\cdot}) = \sum_{g_f=1}^{3} \sum_{g_m=1}^{3} L(z_{i\cdot} \mid g_f, g_m) \mathrm{P}(g_f, g_m) \tag{2.11}$$

Kui vanemad on valitud juhuslikult, siis $\mathrm{P}(g_f, g_m) = \mathrm{P}(g_f) \times \mathrm{P}(g_m)$.

Veelgi enam, kui genotüüpide suhtes kehtib Hardy-Weinbergi seadus, siis on vanemate genotüüpide sagedused üheselt leitavad alleeli *Q* sagedusest *p*:

$$\begin{aligned} &\mathrm{P}(g_f = 1, g_m = 1) = \mathrm{P}(g_f = QQ) \times \mathrm{P}(g_m = QQ) = p^2 \times p^2, \\ &\mathrm{P}(g_f = 2, g_m = 1) = \mathrm{P}(g_f = Qq) \times \mathrm{P}(g_m = QQ) = 2p(1-p) \times p^2, \text{ jne} \end{aligned}$$

Kui genotüüpide arv $n_g > 3$ (põhigeenil on rohkem kui kaks esinemisvormi või määrab tunnuse väärtuse ära enam kui üks geen), on liidetavaid valemis (2.11) lisandunud alleelikombinatsioonide võrra rohkem. Mudelile vastavalt muutuvad ka genotüüpide tõenäosused.

Vanemate fenotüübiväärtuste (z_f, z_m) teadmisel arvestatakse tõepärafunktsiooni kirjapanekul sedagi informatsiooni. Kuna $L(z \mid g) = f(z, \mu_g, \sigma^2)$, siis tõenäosus, et fenotüübiga z isendi genotüüp on g_i, avaldub Bayesi teoreemi alusel seosena

$$\mathrm{P}(g_i \mid z) = \frac{\mathrm{P}(g_i) f(z, \mu_{g_i}, \sigma^2)}{\sum_{j=1}^{n_g} \mathrm{P}(g_j) f(z, \mu_{g_j}, \sigma^2)} = \frac{\mathrm{P}(g_i) f(z, \mu_{g_i}, \sigma^2)}{p(z)},$$

kus $p(z)$ on populatsiooni fenotüübijaotus. Vanemate teadaolevad fenotüübiväärtused võetakse analüüsil arvesse, asendades tõepärafunktsioonis (2.11) genotüübitõenäosused $\mathrm{P}(g)$ tinglike tõenäosustega $\mathrm{P}(g \mid z)$.

Eeldades, et erinevad perekonnad ei ole omavahel suguluses, on kogu andmevektorile **z** vastav üldine tõepärafunktsioon üksikutele perekondadele vastavate tõepärafunktsioonide (2.11) korrutis:

$$L(\mathbf{z}) = \prod_{i=1}^{n_f} L(z_{i\cdot}), \tag{2.12}$$

kus n_f tähistab erinevate perekondade arvu.

Kuigi viimane valem koosneb mitmeist üksteise sees olevaist korrutistest ja summadest, on selles tundmatuid parameetreid vaid viis: kolm genotüübijaotustele vastavat keskväärtust, üks ühine dispersioon σ^2 ja alleelisagedus *p*.

Vältimaks paljude väikese mõjuga geenide ja/või keskkonna toime pidamist tegelikkuses mitteeksisteeriva põhigeeni efektiks, soovitasid R. Elston jt 1975. aastal[1] käsitleda tundmatute parameetritena ka alleelide lahknemistõenäosusi $\mathrm{P}(g_o \mid g_f, g_m)$, mis Mendeli lahknemisseaduste alusel siiani fikseerituiks olid loetud (vt valemid (2.8)).

Tähistame τ_x-ga tõenäosust, et vanema genotüübi x korral saab järglane päranduseks alleeli Q. Dialleelse lookuse korral tuleb siis täiendavalt hinnata kolm τ väärtust, iga genotüübi jaoks üks. Lahknemistõenäosused avalduvad τ definitsioonist lähtudes seostena

$$\begin{aligned}
&\mathrm{P}(qq \mid g_f, g_m) = (1-\tau_{g_f})(1-\tau_{g_m}),\\
&\mathrm{P}(Qq \mid g_f, g_m) = \tau_{g_f}(1-\tau_{g_m}) + \tau_{g_m}(1-\tau_{g_f}),\\
&\mathrm{P}(QQ \mid g_f, g_m) = \tau_{g_f}\tau_{g_m}.
\end{aligned}$$

Valemitega (2.8) antud lahknemistõenäosused saavad τ abil kuju

$$\begin{aligned}
&\mathrm{P}(qq \mid g_f = QQ, g_m = Qq) = (1-\tau_{QQ})(1-\tau_{Qq}),\\
&\mathrm{P}(Qq \mid g_f = QQ, g_m = Qq) = \tau_{QQ}(1-\tau_{Qq}) + \tau_{Qq}(1-\tau_{QQ}),\\
&\mathrm{P}(QQ \mid g_f = QQ, g_m = Qq) = \tau_{QQ}\tau_{Qq}
\end{aligned}$$

ja tõepärafunktsioon (2.9) selle konkreetse perekonna kohta väljakirjutatuna on

$$\begin{aligned}
L(z_{ij} \mid QQ, Qq) = {} & \tau_{QQ}\tau_{Qq} \cdot f(z_{ij}, \mu_{QQ}, \sigma^2)\\
& + \left[\tau_{QQ}(1-\tau_{Qq}) + \tau_{Qq}(1-\tau_{QQ})\right] \cdot f(z_{ij}, \mu_{Qq}, \sigma^2)\,.\\
& + (1-\tau_{QQ})(1-\tau_{Qq}) \cdot f(z_{ij}, \mu_{qq}, \sigma^2)
\end{aligned}$$

On lihtne näha, et see tõepärafunktsioon võrdub Mendeli lahknemisseaduste alusel arvutatud tõenäosuste $\tau_{QQ} = 1$ ja $\tau_{Qq} = 1/2$ korral tõepärafunktsiooniga (2.9).

Põhigeeni hüpoteesi tõestatuks lugemisel peavad olema täidetud kolm tingimust: (1) võrreldes tavalise normaaljaotusega kirjeldab segujaotus andmeid oluliselt paremini, (2) ei õnnestu kummutada hüpoteesi alleelide lahknemisest vastavalt Mendeli lahknemisseadustele ($\tau_{QQ} = 1$, $\tau_{Qq} = 1/2$ ja $\tau_{qq} = 0$), (3) kummutatakse hüpotees kõigi lahknemistõenäosuste võrdsusest ($\tau_{QQ} = \tau_{Qq} = \tau_{qq}$). Kaks lisatingimust hoolitsevad selle eest, et iga fenotüübijaotuse kõrvalekallet normaaljaotusest või indiviidide grupile mõjuva ühtse keskkonna efekti ei peetaks põhigeeni mõjuks.

Näide. McGuffin, P., P. Huckle. 1990. Simulation of mendelism revisited: the recessive genefor attending medical school. *Am. J. Hum. Genet.* 46: 994-999.

Uuritavaks tunnuseks on meditsiinikoolis käimine väärtustega

$$z = \begin{cases} 1, & \text{käis meditsiinikoolis} \\ 0, & \text{ei käinud meditsiinikoolis} \end{cases}.$$

Vaatluse all olnud Walesi Meditsiinikollegi 249-st tudengist oli 13,4%-l kas isa või ema samuti kunagi meditsiinikoolis käinud.

Olgu populatsiooni keskmine väärtus μ. Erinevatele genotüüpidele vastavad keskmised erinevused sellest on põhigeeni olemasolul esitatavad otsitavas lookuses paiknevate alleelide aditiivse mõju a, dominantsusest (alleelide lookusesisesest interaktsioonist) tingitud efekti d ning alleelisageduse p abil. Segregatsioonanalüüsiga võrreldi üldist põhigeenieeldusel konstrueeritud mudelit (hinnatavaiks parameetriteks a, d ja p) ja retsessiivset mudelit (hinnatavaiks parameetriteks a ja p, $d = 0$) nn nullmudeliga ($a = d = p = 0$). Lisaks võrreldi Mendeli lahknemistõenäosustel ($\tau_1 = 1$, $\tau_2 = 1/2$ ja $\tau_3 = 0$) baseeruvat mudelit kahe alternatiivse mudeliga: ühel juhul hinnati tõenäosused τ_1, τ_2 ja τ_3 andmete põhjal ja teisel juhul eeldati nende võrdsust ($\tau_1 = \tau_2 = \tau_3$ – segumudeli liidetavaid arvestati võrdsete kaaludega).

Erinevatele mudelitele vastavad logaritmilise tõepärafunktsiooni väärtused on toodud näite lõpus olevas tabelis.

Üldine põhigeeni eeldusel konstrueeritud mudel lähendab andmeid tunduvalt paremini kui nullmudel, tõepärasuhte statistiku väärtuseks tuleb $283{,}60 - 120{,}14 = 163{,}46$ (3 vabadusastet). Et aga retsessiivne põhigeeni mudel võimaldab väiksema parameetrite arvu

[1] Elston, R. C., K. K. Nasmboodiri, C. J. Glueck, R. Fallat, R. Tsang ja V. Leuba. 1975. Studies of the genetic transmission of hypercholesterolemia and hypertriglyceridemia in a 195 member kindred. *Ann. Hum. Genet.* **39**: 67-87

juures sama häid prognoose kui üldine põhigeeni mudel (põhigeeni alleelide lookusesisese interaktsiooni arvestamine ei suurenda mudeli kirjeldavuse määra), on edasistes võrdlustes aluseks võetud retsessiivne põhigeeni mudel. Võrreldes alleelide lahknemise seaduspärasid mittearvestava mudeliga (võrdsed lahknemistõenäosused) kirjeldab retsessiivne mudel andmeid märksa paremini. Sellega on kaks Elstoni poolt välja pakutud tingimustest rahuldatud – (1) jaotuste segu sobitub andmetele paremini kui tavaline normaaljaotus ja (3) väide lahknemistõenäosuste võrdsusest on ümber lükatud. Kui nüüd aga võrrelda omavahel ühe dialleelse põhigeeni eeldusel fikseeritud lahknemistõenäosustega ($\tau_1 = 1$, $\tau_2 = 1/2$ ja $\tau_3 = 0$) retsessiivset mudelit ja andmetest hinnatud lahknemistõenäosustega mudelit, selgub, et viimane lähendab andmeid oluliselt paremini, tõepärasuhte statistiku väärtuseks tuleb $120{,}14 - 111{,}22 = 8{,}92$ (3 vabadusastet, $p < 0{,}03$). Seega ei vasta uuritavad andmed teisele põhigeeni olemasolu kriteeriumile – hüpotees alleelide Mendeli seaduste järgsest lahknemisest on ümber lükatud. Meditsiiniõpingutega alustamist soosib ilmselt hoopiski ümbritsev keskkond ja vanemate mõju (perekonnatraditsioonid jne).

Mudel	Parameetrid (lisaks üldkeskmisele μ)		const + −2 ln (tõepäraf)
	Hinnatavad	Fikseeritud	
Nullmudel (geneetilised efektid puuduvad)	–	$a = d = p = 0$	283,60
Üldine põhigeeni mudel	a, d, p	$\tau_1 = 1$, $\tau_2 = 1/2$, $\tau_3 = 0$	120,14
Dominantsiefektita põhigeeni mudel	a, p	$d = 0$, $\tau_1 = 1$, $\tau_2 = 1/2$, $\tau_3 = 0$	120,14
Võrdsed lahknemis-tõenäosused	a, p, $\tau_1 = \tau_2 = \tau_3$	$d = 0$	283,60
Empiirilised lahk-nemistõenäosused	a, p, τ_1, τ_2, τ_3	$d = 0$	111,22

¤ Ühise perekonna efektid

Ühe perekonna järglastele mõjub keskkond enamasti sarnaselt. Selle arvestamiseks tuleb eelnevalt tuletatud tõepärafunktsioone vähe modifitseerida. Tähistame *i*. perekonna efekti c_i ning eeldame, et $c_i \sim N(0, \sigma_c^2)$. Nagu varemgi eeldame, et igale genotüübile vastavad fenotüübiväärtused jaotuvad (tinglikuna ühise perekonnaefekti c_i suhtes) normaalselt, seejuures avaldub perekonna *i* genotüübiga g_o järglase keskmine fenotüübiväärtus summana $\mu_{g_o} + c_i$, fenotüübiväärtuste varieeruvust kirjeldab endiselt dispersiooniparameeter σ^2.

Perekonna *i*, kus on n_i järglast, jaoks väljakirjutatud tõepärafunktsioon avaldub korrutisena

$$L(z_{i\cdot} \mid g_f, g_m, c_i) = \prod_{j=1}^{n_i} \left[\sum_{g_{oj}=1}^{3} \mathrm{P}(g_{oj} \mid g_f, g_m) \times f(z_{ij}, \mu_{g_{oj}} + c_i, \sigma^2) \right]. \qquad (2.13)$$

Keskmine tõepärafunktsiooni väärtus üle kõigi võimalike ühise perekonna efekti c_i väärtuste avaldub integraalina

$$L(z_{i\cdot} \mid g_f, g_m) = \int_{-\infty}^{\infty} L(z_{i\cdot} \mid g_f, g_m, c) \times f(c, 0, \sigma_c^2)\, dc\,. \qquad (2.14)$$

Summeerides tõepärafunktsiooni (2.14) üle vanemate kõigi võimalike genotüüpide paaride (põhigeeni suhtes), saame analoogselt avaldisele (2.11) tingimatu tõepärafunktsiooni perekonna *i* suhtes. Kui genotüübid vastavad Hardy-Weinbergi tasakaalu tingimustele, sisaldab tingimatu tõepärafunktsioon kuut hinnatavat parameetrit: alleelisagedus *p*, kolm genotüübijaotustele vastavat keskväärtust ja dispersioonid σ^2 ning σ_c^2. Tingimatute tõepärafunktsioonide korrutis üle kõigi perekondade annab meile analoogselt avaldisega (2.12) üldise tõepärafunktsiooni.

Kitsendatud mudelile, kus eeldatakse küll ühise perekonna mõju, aga mitte põhigeeni olemasolu, vastav *i*. perekonna jaoks kirja pandud tõepärafunktsioon on

$$L(z_{i\cdot}) = \int_{-\infty}^{\infty} L(z_{i\cdot} \mid c) \times f(c, 0, \sigma_c^2)\, dc = \int_{-\infty}^{\infty} \left[\prod_{j=1}^{n_i} f(z_{ij}, \mu + c, \sigma^2) \right] \times f(c, 0, \sigma_c^2)\, dc\,. \qquad (2.15)$$

Testimaks ühise perekonnaefekti olemasolu põhigeeni puudumisel sobib tõepärasuhte statistik, mis vastandab võrduse (2.15) juhu $\sigma_c^2 = 0$ jaoks genereeritud tõepärafunktsioonile. See viimane kujutab enesest aga tavalist normaaljaotust, millele vastava tõepärafunktsiooni maksimum on defineeritud valemiga (2.6).

Tõepärasuhte statistik, kontrollimaks küll põhigeeni, aga mitte ühise perekonnaefekti olemasolu, võrdleb omavahel täielikule mudelile vastavat ja eeldusel $\sigma_c^2 = 0$ genereeritud tõepärafunktsiooni.

¤ Polügeenne taust

Kui statistiku käest küsida, mida tähendab termin **segamudel**, vastab ta üsna suure tõenäosusega, et see tähendab nii fikseeritud kui ka juhuslike efektidega üldist lineaarset mudelit. Geneetikule sarnast küsimust esitades saame aga enamasti vastuseks, et see on põhigeeni pluss polügeense taustaefekti vanemailt järglastele edasikandumist kirjeldav mudel. Ja õigus on neil mõlemal, sest mõiste kasutamise traditsioonid on lihtsalt erinevad. Selles punktis anname lühidalt ülevaate põhigeeni ja polügeense efektiga segregatsioonanalüüsi segumudelist, ehk siis **segregatsioonanalüüsi segamudelist**, nagu seda geneetikaalases kirjanduses tihti nimetatakse. Konspektis edaspidi tähendab termin segamudel aga ikkagi üldise lineaarse mudeli levinud erijuhtu.

Eeldame, et suure hulga väikese mõjuge geenide summaarne efekt A on normaaljaotusega $N(0, \sigma_A^2)$. Genotüübiga g ja polügeense efektiga A indiviidi fenotüübiväärtus on samuti on normaaljaotusega keskväärtusega $\mu_g + A$ ja dispersiooniga σ_E^2. Jättes praegu arvestamata võimaliku ühise perekonna mõju, on tinglik tõepärafunktsioon i. perekonna j. järglase jaoks esitatav kujul

$$L(z_{ij} \mid g_f, g_m, A_{oj}) = \sum_{g_{oj}=1}^{3} \mathrm{P}(g_{oj} \mid g_f, g_m) \cdot f(z_{ij}, \mu_{g_{oj}} + A_{oj}, \sigma_E^2),$$

kus g_f ja g_m märgivad jällegi vastavalt isa ja ema genotüüpi ning A_{o_j} on j. järglasele omane polügeenne efekt. Tõepärafunktsiooni tinglikkus paljude geenide aditiivse efekti A_o suhtes kõrvaldatakse kahes osas. Esmalt asendatakse see tinglikkusega vanemate polügeensete efektide (A_f, A_m) suhtes:

$$L(z_{ij} \mid g_f, g_m, A_j, A_m) = \int_{-\infty}^{\infty} L(z_{ij} \mid g_f, g_m, A_{oj}) p(A_{oj} \mid A_f, A_m) dA_{oj}. \qquad (2.16)$$

Et vanemate polügeensed efektid on normaaljaotusega, on seda ka järglase vastav efekt, keskväärtusega $(A_f + A_m)/2$ ja dispersiooniga $\sigma_A^2/2$. Tinglik tihedusfunktsioon avaldises (2.16) on siis asendatav normaaljaotuse tihedusfunktsiooniga:

$$p(A_o \mid A_f, A_m) = f\left(A_o, \frac{A_f + A_m}{2}, \frac{\sigma_A^2}{2}\right).$$

Teiseks, integreerides avaldist (2.16) üle vanemate kõigi võimalike polügeensete taustaefektide, saame vaid otsitava põhigeeni suhtes tingliku tõepärafunktsiooni i. perekonna jaoks

$$L(z_{i\cdot} \mid g_f, g_m) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[\prod_{j=1}^{n_i} L(z_{ij} \mid g_f, g_m, A_f, A_m)\right] f(A_f, 0, \sigma_A^2)\ f(A_m, 0, \sigma_A^2)\ dA_f\ dA_m. \qquad (2.17)$$

Viimane avaldis eeldab, et vanemad on valitud juhuslikult ega ole omavahel sugulased. Selle eelduse mittetäidetuse korral tuleks tõepärafunktsioon (2.16) keskmistada üle A_f ja A_m ühistiheduse. Viimase sammuna summeeritakse avaldis (2.17) üle põhigeeni kõigi alleelikombinatsioonide (vt valem (2.11)) ja saadakse tingimatu tõepärafunktsioon ühe konkreetse perekonna jaoks. Hinnatavaid parameetreid on selles kuus: kolm põhigeeni genotüüpidele vastavat keskväärtust, alleeli sagedus p (eeldades Hardy-Weinbergi tasakaalu) ning dispersiooniparameetrid σ_E^2 ja σ_A^2.

Põhigeeni olemasolu kontrollimiseks võrreldakse eelnevalt kirjeldatud mudelit kitsendatud mudeliga, mis eeldab küll polügeense taustaefekti, aga mitte põhigeeni ega ühise keskkonna (perekonna) mõju olemasolu. Konkreetse indiviidi jaoks välja kirjutatud tema vanemate polügeensete efektide suhtes tinglik tõepärafunktsioon on

$$L(z_{ij} \mid A_f, A_m) = \int_{-\infty}^{\infty} f(z_{ij}, \mu + A_o, \sigma_E^2)\ f\left(A_o, \frac{A_f + A_m}{2}, \frac{\sigma_A^2}{2}\right) dA_o,$$

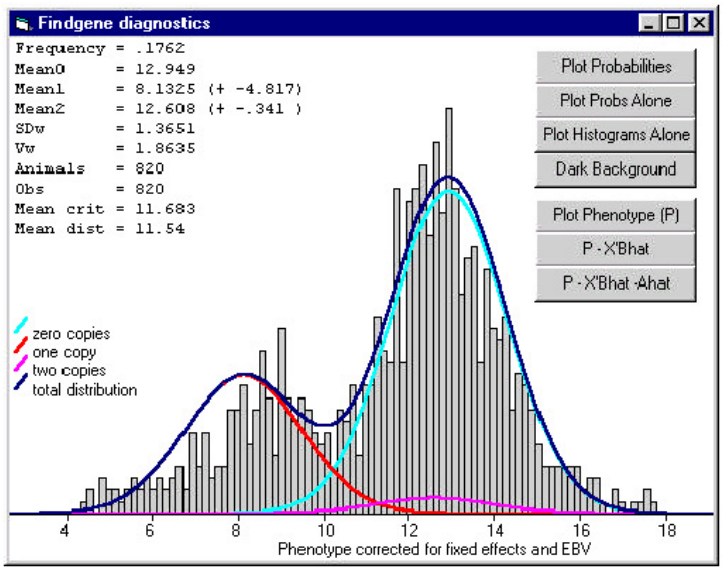
millest tingimatu tõepärafunktsiooni saame jällegi üle vanemate kõigi võimalike polügeensete taustaefektide integreerides:

$$L(z_{i\cdot}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[\prod_{j=1}^{n_i} L(z_{ij} \mid A_f, A_m)\right] f(A_f, 0, \sigma_A^2)\ f(A_m, 0, \sigma_A^2)\ dA_f\ dA_m. \qquad (2.18)$$

Ühise keskkonna mõju lisamine avaldistesse (2.17) või (2.18) peaks juba lihtne olema.

¤ Keerukamad mudelid

Et suuremahuliste praktiliste uuringute korral osutuvad eelnevalt kirjeldatud segregatsioonanalüüsi mudelid liialt lihtsameelseteks, on välja töötatud mudelid ka märksa keerukamate juhtude tarvis – segregatsioonanalüüsil saab arvesse võtta kõiki võimalikke uuritavate indiviidide vahelisi sugulussidemeid, tõepärafunktsioonid võib konstrueerida arvestamaks ükskõik kui paljude fikseeritud faktoritega (vanus, sugu, spetsiifiline keskkond), analüüsida on võimalik ka mitmemõõtmelisi mudeleid. Ja idee on kõigi nende mudelite korral sama – esmalt tinglikustamine kõigi võimalike genotüüpide suhtes ja seejärel keskmistamine, et saada tingimatuid tõepärafunktsioone. Reaalsete arvutuste tarvis on välja töötatud mitmeid ligikaudseid meetodeid ning nende rakendamiseks arvutiprogramme (joonis 13). Enamasti kasutatavate iteratiivsete meetodite (vt näiteks Stricker jt[2]) kõrval arenesid 20. sajandi viimasel kümnendil jõudsalt ka taasvaliku meetodid, millest tuntuim on kindlasti Gibbsi valik (vt näiteks Janss jt[3]). Kõigi nende korral antakse esmalt ette mingi eelinformatsioon (alleelisagedus, erinevate genotüüpide efektid), mida järgnevalt samm-sammult reaalsete põlvnemisandmete alusel parandatakse, teise etapina viiakse hinnatud genotüübisageduste baasil läbi dispersioonanalüüsi segamudelite analüüs, võtmaks arvesse polügeenset efekti ja keskkonna mõju, seejärel liigutakse uuesti esimese etapi juurde, kus korrigeeritud fenotüübiandmete jaotust püütakse lähendada põhigeeni eeldusel konstrueeritud jaotuste seguga, jne, kuni protseduur koondub.



Joonis 13. Programmi *Findgene* tulemuste aken – uuritavate loomade fenotüübijaotus on korrigeeritud fikseeritud efektide ja polügeense mõju (aretusväärtuste) suhtes ning esitatud seejärel dialleelse põhigeeni kolmele genotüübile vastavate jaotuste seguna.

2.4.3 Diskreetsete tunnuste analüüs

Normaaljaotuse eeldustel tuletatud tõepärafunktsioonid on kergelt modifitseeritavad teostamaks komplekssed segregatsioonanalüüsi ka diskreetsete tunnuste, nagu näiteks haiguse esinemine või mitte esinemine, korral. Defineerime genotüübile g vastava **penetrantsuse** ψ_g kui tõenäosuse, et uuritav tunnus avaldub juhuslikult valitud indiviidil genotüübiga g. Kodeerides uuritava tunnuse järgmiselt:

$$y = \begin{cases} 0, & \text{tunnus ei avaldunud} \\ 1, & \text{tunnus avaldus} \end{cases},$$

saame indiviidile genotüübiga g vastava tõepärafunktsiooni esitada kujul

$$L(y \mid g) = (\psi_g)^y (1-\psi_g)^{1-y} = \begin{cases} 1-\psi_g, & \text{kui } y = 0 \\ \psi_g, & \text{kui } y = 1 \end{cases}. \qquad (2.19)$$

[2] Stricker, C., R. L. Fernando, R. S. Elston. 1995. An algorithm to approximate the likelihood for pedigree data with loops by cutting. *Theor. Appl. Genet.* 91: 1054-1063.

[3] Janss, L. L. G., R. Thompson, J. A. M. Van Arendonk. 1995. Applications of Gibbs sampling for inference in mixed model gene-polygenic inheritance model in animal populations. *Theor. Appl. Genet.* 91: 1137-1147.

Üldisemalt, kui tunnusel on n erinevat (diskreetset) väärtust ja $\psi_{k,g}$ märgib tõenäosust, et indiviidil genotüübiga g on fenotüübis avaldunud tunnuse väärtus k, esitub tõepärafunktsioon kujul

$$L(y \mid g) = \prod_{k=1}^{n} (\psi_{k,g})^{\delta(y,k)}, \text{ kus } \delta(y,k) = \begin{cases} 1, & \text{kui } y = k \\ 0, & \text{ülejäänud juhtudel} \end{cases}.$$

¤ **Ühelookuseline mudel**

Oletame, et binaarse tunnuse väärtused määrab ära üks dialleelne lookus, ning tähistame genotüüpide QQ, Qq ja qq penetrantsust vastavalt ψ_1, ψ_2 ja ψ_3. Kui alleeli Q sagedus on p, siis on uuritava tunnuse esinemise tõenäosus Hardy-Weinbergi tasakaalus olevas populatsioonis

$$K = p^2 \times \psi_1 + 2p(1-p) \times \psi_2 + (1-p)^2 \times \psi_3 . \tag{2.20}$$

Vajalikud tõepärafunktsioonid konstrueeritakse seosest (2.19), tinglikustades selle uuritavate isendite vanemate genotüüpide suhtes. Olgu meil näiteks vaatluse all olevas i. perekonnas n järglast, fenotüübiväärtustega uuritava tunnuse osas $y_{i1}, \ldots, y_{in}$. Selle perekonna j. järglase jaoks kirja pandud vanemate genotüüpide suhtes tinglik tõepärafunktsioon on

$$L(y_{ij} \mid g_f, g_m) = \sum_{g_o=1}^{3} \mathrm{P}(g_o \mid g_f, g_m)\, (\psi_{g_o})^{y_{ij}} (1-\psi_{g_o})^{1-y_{ij}} .$$

Korrutades viimase avaldise üle kõigi järglaste ning summeerides seejärel üle vanemate kõigi võimalike genotüüpide, saame analüüsiks vajaliku tingimatu tõepärafunktsiooni, mis sisaldab nelja tundmatut parameetrit: alleeli sagedust p ja kolme tõenäosust ψ_i. Juhul, kui lugeda teadaolevaks ka tunnuse esinemise suhteline sagedus (2.20), kahaneb hinnatavate parameetrite arv ühe võrra.

2.4.4 Segumudeli parameetrite hindamine EM-meetodil

EM-algoritmi idee põhineb tõdemusel, et mõnede vaatluste kohta enama informatsiooni omamisel on parameetrite suurima tõepära hinnangud kergelt leitavad. Näiteks teades uuritavate indiviidide genotüüpe, saame genotüübijaotuste parameetrite hinnangud vastavate genotüüpidega indiviidide kogumeid analüüsides. Algoritmi sissejuhatava sammuna antakse ette parameetrite alglähendid. Nendele vastavalt leitakse esimese, nö E-sammuna (E = *expectation*), oodatav lisainformatsioon (indiviidide mingisse genotüübiklassi kuuluvus). Viimase alusel arvutatakse hinnatavate parameetrite suurima tõepära hinnangud (nö M-samm, M = *maximization*). Saadud hinnangutega pöördutakse tagasi E-sammu juurde jne.

Kujutagu uuritava tunnuse z fenotüübijaotus $p(z)$ enesest kaalutud genotüübijaotuste $f(z, \mu_k, \sigma^2)$, k = 1,…,m, kus kaaluparameetriteks on genotüübitõenäosused $\pi_k = \mathrm{P}(k)$, summat. Tähistame $w(k \mid z) = \mathrm{P}(k \mid z)$ tõenäosust, et indiviid fenotüübiväärtusega z omab genotüüpi k. Bayesi teoreemi abil on see tinglik tõenäosus (konkreetsele indiviidile vastav kaaluparameeter) esitatav suhtena

$$w(k \mid z) = \frac{\mathrm{P}(k) \times \mathrm{P}(z \mid k)}{\mathrm{P}(z)} = \frac{\pi_k \times f(z, \mu_k, \sigma^2)}{\mathrm{P}(z)} = \frac{\pi_k \times f(z, \mu_k, \sigma^2)}{\sum_{j=1}^{m} \pi_j \times f(z, \mu_j, \sigma^2)} .$$

Selliselt püstitatud mudeli parameetrite hindamine EM-meetodil toimub järgmiste sammudena:

Sissejuhatav samm. Valitakse alglähendid hinnatavale dispersioonile $\hat{\sigma}^{2(0)}$, keskväärtustele $\hat{\boldsymbol{\mu}}^{(0)} = (\hat{\mu}_1^{(0)}, \ldots, \hat{\mu}_m^{(0)})$ ja seguproportsioonidele $\hat{\boldsymbol{\pi}}^{(0)} = (\hat{\pi}_1^{(0)}, \ldots, \hat{\pi}_m^{(0)})$.

(1) **E-samm**. Hinnanguid $\hat{\sigma}^{2(0)}$, $\hat{\boldsymbol{\mu}}^{(0)}$ ja $\hat{\boldsymbol{\pi}}^{(0)}$ kasutades, leitakse Bayesi valemist tõenäosus, et indiviid fenotüübiväärtusega z_i kuulub genotüübiklassi k:

$$w^{(1)}(k \mid z_i) = \frac{\hat{\pi}_k^{(0)} \times f(z_i, \hat{\mu}_k^{(0)}, \hat{\sigma}^{2(0)})}{\sum_{j=1}^{m} \left[\hat{\pi}_j^{(0)} \times f(z_i, \hat{\mu}_j^{(0)}, \hat{\sigma}^{2(0)}) \right]},$$

kus $f(z_i, \hat{\mu}_k^{(0)}, \hat{\sigma}^{2(0)})$ märgib normaaljaotust keskväärtusega $\hat{\mu}_k^{(0)}$ ja dispersiooniga $\hat{\sigma}^{2(0)}$ kohal z_i (i = 1, …, n; n – indiviidide arv).

(2) **M-samm**. Kaale $w^{(1)}(k \mid z_i)$ kasutades leitakse uued hinnangud segumudeli parameetritele:

(a) segu proportsioonid leitakse kui keskmised tõenäosused kuuluda klassi k:

$$\hat{\pi}_k^{(1)} = \bar{w}_k^{(1)} = \frac{1}{n} \sum_{i=1}^{n} w^{(1)}(k \mid z_i) ;$$

(b) keskväärtused leitakse vaatluste kaalutud keskmistena:

$$\hat{\mu}_k^{(l)} = \frac{1}{n}\sum_{i=1}^{n} z_i \left(\frac{w^{(l)}(k \mid z_i)}{\bar{w}_k^{(l)}} \right);$$

(c) dispersioonid leitakse vaatluste kaalutud dispersioonidena:

$$\hat{\sigma}^{2(l)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m} \left(z_i - \hat{\mu}_k^{(l)} \right)^2 w^{(l)}(k \mid z_i) .$$

M-sammul arvutatud parameetrite hinnangutega pöördutakse uuesti E-sammu juurde, et leida täpsustatud kaalud kõigile indiviididele, jne, kuni protseduur koondub.

2.5 Indiviidide genotüüpide hindamine

Peale põhigeeni leidmist on loomulik püüda välja selgitada ka üksikute indiviidide genotüübid. Näiteks kui tunnuse väärtus on määratud korraga nii põhigeeniga kui ka polügeense efektiga, võivad mõned indiviidid erineda teistest põhigeeni genotüübi poolest, mõned aga polügeense mõju poolest. Üksikindiviidide genotüüpide teadmine aitab täpsustada molekulaarsel tasandil uuritavate isendite hulka.

Matemaatiliselt põhineb üksikute indiviidide genotüüpide hindamine Bayesi teoreemil, mille kohaselt avaldub tõenäosus, et indiviid fenotüübiväärtusega z omab genotüüpi j ($1 \le j \le m$), võrdusena

$$\mathrm{P}(g_j \mid z) = \frac{\mathrm{P}(g_j)\,\mathrm{P}(z \mid g_j)}{\mathrm{P}(z)} = \frac{\mathrm{P}(g_j)\,\mathrm{P}(z \mid g_j)}{\sum_{i=1}^{m} \mathrm{P}(g_i)\,\mathrm{P}(z \mid g_i)} .$$

Erinevate alleelide pärandumise jälgimiseks ja isendite genotüüpide täpsemaks määramiseks on loomulik kaasata uuringutesse võimalikult põhjalikud andmed põlvnemise ja sugulaste kohta. Kogu selle info optimaalne ärakasutamine pole aga enam sugugi nii lihtne.

Huvilised võivad selle segregatsioonanalüüsi valdkonnaga lähemalt tutvuda järgmiste artiklite vahendusel.

van Arendonk, J. A. M., C. Smith, B. W. Kennedy. 1989. Method to estimate genotype probabilities at individual loci in farm livestock. *Theor. Appl. Genet.* 78: 735-740.

Kerr, R. J., B. P. Kinghorn. 1996. An efficient algorithm for segregation analysis in large populations. *J. Anim. Breed. Genet.* 113: 457-469.

Kinghorn, B. P. 1997. An index of information content for genotype probabilities derived from segregation analysis. *Genetics* 145: 479-483.

NB! Loomade aretuses mõistetakse segregatsioonanalüüsi all sageli just indiviidide genotüüpide hindamist.

Ülesanne 2.

Olgu meil vaatluse all järgmine põlvnemisskeem, kus kõigi emaste isendite genotüüp ühe dialleelse lookuse suhtes on määratud. Leidke, milliseid alleele ja millise tõenäosusega kannavad skeemil küsimärkidega tähistatud isased indiviidid. Alleeli A sagedus on eelnevate uuringute alusel 0,6.