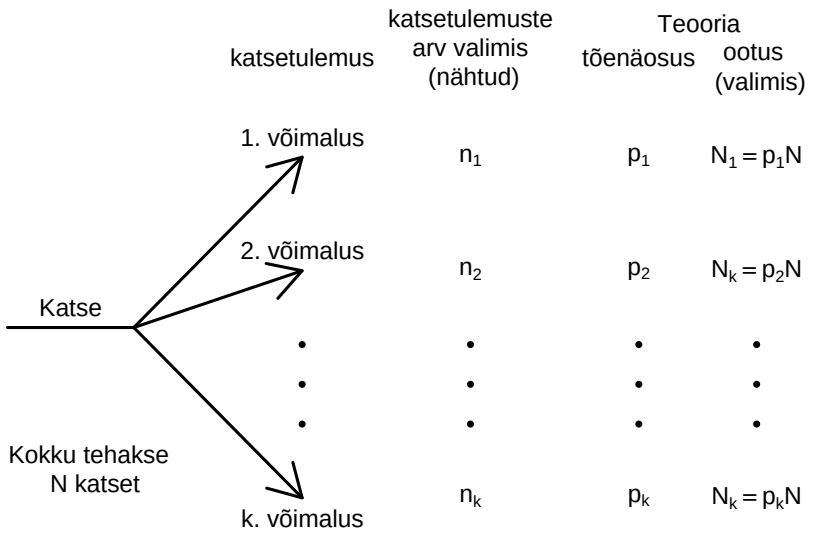
7.7 Hii-ruut test

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 7.7.

Joonis 7.3: Tegelikud katsetulemused ja teooria ennustus



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

Näide 7.6 *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 7.3.

Tabel 7.3: Mendeli seaduste paikapidavuse kontrollimine

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

Näide 7.7 *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 7.4.*

Tabel 7.4: 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 7.3 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 7.4 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 7.5.

Näide 7.8 *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 7.5: 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 7.6: 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 7.7.*

Tabel 7.7: 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.

7.7.1 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).

7.7.2 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 7.9 *Vaatame, kas saame tõestada, et tudengite alkoholitarbimise ja nende soo vahel eksisteerib seos. Algandmed on esitatud tabelis 7.8.*

Esmalt hindame nii õlletarbimise kui ka soolise jaotuse (tabelid 7.9 ja 7.10).

Nüüd leiame tõenäosused $P(sugu = i, õlu = j)$. Leitud tõenäosused on toodud tabelis 7.11.

Nüüd leiame teooria (sõltumatus) poolt ennustatavad tudengite arvud $N_{sugu=i,õlu=j}$, vaata tabel 7.12.

Tabel 7.8: 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 7.9: 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 7.10: Sooline jaotus

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

Tabel 7.11: 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 7.12: Sõltumatuse korral oodatavad tudengite arvud

sugu|õlu	ei tarbi	alla pudeli	1-4	5 või enam	kokku
naine	0.3117*611=206.04	206.04	71.26	28.66	512
mees	59.96	59.96	20.74	8.34	149
kokku	266	266	92	37	611

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.*

7.8 Ülesanded

1. Ristati haplotüüpidega Aa, Bb ja Aa, Bb isendeid. Saadud andmed on toodud tabelis 7.13. 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 7.13: 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!)