
Konpekt aines

MITTEPARAMEETRILINE STATISTIKA

2017 sügissemester

KONSPEKTI AUTOR: MÄRT MÖLS

Peatükk 1

Sissejuhatus

1.1 Mitteparameetrilise statistika mõistest

Mitteparameetrilise statistika mõiste on lai ja kaks mitteparameetrilisele statistikale pühendatud raamatut võivad käsitleda (esmapilgul) täiesti erinevaid valdkondi. Sestap alustuseks paari sõnaga sellest, mida mõeldakse mitteparameetrilise statistika ja mitteparameetriliste meetodite all.

Klassikalises statistikas nõutakse sageli, et me teaksime uuritava tunnuse kohta eelinformatsiooni — näiteks seda, millisesse jaotuste perre ta kuulub (kas tegemist on normaaljaotusega, eksponentjaotusega, gammajaotusega või hoopis ...-jaotusega). Teades jaotuste peret (näiteks normaaljaotus) piisab vaid paari parameetri (keskväärtuse, dispersiooni) hindamisest uuritava tunnuse jaotuse üheseks määramiseks. Teades neid paari parameetrit, saame vastata mistahes küsimusele uuritava tunnuse jaotuse kohta. Sestap pöörabki klassikaline parameetriline statistika suurt tähelepanu jaotusparameetrite hindamisele ja nende kohta hüpoteeside kontrollimisele.

Paraku praktikas osutub kitsa parameetrilise jaotuste pere täpne määramine sageli vägagi raskeks. Seega praktiliste ülesannete lahendamisel puututakse sageli kokku nn mitteparameetriliste jaotuste peredega.

Definitsioon 1.1 *Jaotuste peret kutsutakse **mitteparameetriliseks jaotuste perek**s, kui vastavasse perre kuuluvat jaotust ei saa üheselt identifitseerida kasutades vaid vähest arvu parameetreid (üheks selliseks jaotuste perek on näiteks kõigi pidevate jaotuste pere).*

Statistilist meetodit, mis on rakendatav tunnus(t)e uurimiseks, mille jaotus kuulub mitteparameetriliste jaotuste perre, nimetatakse **mitteparameetriliseks meetodiks**. Juhul, kui tahetakse rõhutada, et üks või teine meetod

ei eelda mõnda lihtsat jaotust (nagu normaaljaotust või beetajaotust) siis räägitakse ka jaotusvabadest (*distribution-free*) meetoditest.

Vaatame ühte teist näidet. Tunneme huvi, milline on tunnuse Y tinglik jaotus juhul, kui teame tunnust X , $f_{Y|X}$. Oletame, et Y -tunnuse tinglikuks jaotuseks on normaaljaotus. Tavapärane, parameetriline lähenemine nõuab, et teaksime tunnuste Y ja X vahelist seost piisavalt täpselt, et kõigest peale mõne täiendava parameetri hindamist saame teada seose tunnuste vahel. Näiteks eeldame, et $Y|X \sim N(c_0 + c_1X, \sigma^2)$. Niipea kui saame teada parameetrite c_0 ja c_1 väärtused (või hindame need), teame ka seost tunnuste vahel.

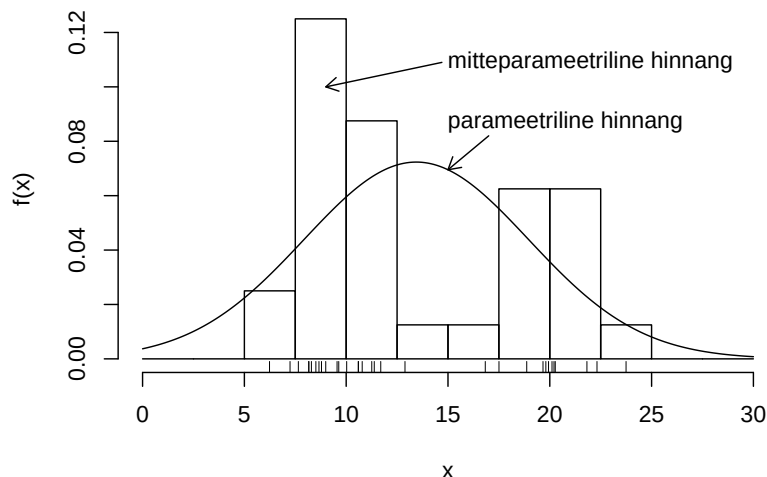
Vahel pole aga seost tunnuste vahel võimalik piisavalt täpselt eelnevalt kirja panna. Mõnikord saame kõigest öelda, et $Y|X \sim N(f(X), \sigma^2)$, kus $f()$ on näiteks mingi suvaline pidev funktsioon. Sellisel juhul pole võimalik määratleda uuritava tunnuse (tinglikku) jaotust vaid paari tundmatu parameetri hindamise abil. Sestap kutsutakse ka toodud näites kahe tunnuse vahelise seose leidmist mitteparameetriliseks regressiooniks. Kui soovitakse rõhutada, et mõned eeldused siiski on tehtud (normaaljaotus), aga midagi pole suudetud/osatud/tahetud parametrizeerida (funktsiooni $f()$), siis räägitakse ka semiparameetrilisest (*semiparametric*) või poolparameetrilisest meetodist.

Järgnevalt vaatame mõningaid näiteid, kuidas erinevatele küsimustele saab leida vastuseid mitteparameetrilisi meetodeid kasutades.

1.2 Tihedusfunktsiooni hindamine

Tihedusfunktsiooni hindamisest räägime pikemalt hiljem. Siin toome lihtsalt ühe näite, rõhutamaks parameetrilise ja mitteparameetrilise statistika erinevust. Parameetrilise statistika puhul eeldatakse, et vaatlused on mingist jaotuste perest (näiteks eeldatakse, et tegemist on normaaljaotusega), hinnatakse parameetrid — keskväärtust hinnatakse näiteks valimi keskmisega ja populatsiooni dispersiooni saab hinnata valimi dispersiooniga ning ongi meil leitud hinnang uuritava tunnuse tihedusfunktsioonile. Mitteparameetrilise meetodi näitena võib aga tuua histogrammi. Vaata näiteks joonist 1.1.

Joonis 1.1: Hinnang tihedusfunktsioonile



1.3 Märgitest

Paljud testid (t -test ja regressioon- ning dispersioonanalüüsis kasutatav F -test, tõepärasuhtel baseeruvad testid jne) eeldavad, et me teaksime uuritava tunnuse jaotust. Vaatame, kuidas on võimalik hüpoteese kontrollida eeldamata uuritava tunnuse jaotuse kohta midagi peale pidevuse.

Märgitest võimaldab kontrollida hüpoteese uuritava tunnuse mediaani kohta,

$$H_0 : \text{med}(X) = \Theta_0$$

$$H_1 : \text{med}(X) \neq \Theta_0.$$

Kuidas konstrueerida sobivat teststatistikut? Mediaan on väärtus, millest väiksemaid ja suuremaid väärtuseid peaks olema ühepalju. Seega, kui Θ_0 on tegelikult mediaan, peaks juhuslikest suurustest $X_1 - \Theta_0, \dots, X_n - \Theta_0$ ligikaudu pooled olema suuremad nullist (vaatlus oli suurem mediaanist) ja ligikaudu pooled väiksemad nullist (vaatlus oli väiksem mediaanist). Täpsemalt öeldes:

$$P(X_i - \Theta_0 > 0) = P(X_i - \Theta_0 < 0) = 1/2,$$

kui Θ_0 on ka tegelikult mediaan. Meie teststatistik võiks loendada, kui mitmel korral nägime valimis Θ_0 -st suuremat väärtust,

$$B(X, \Theta_0) = \sum_{i=1}^n I_{X_i - \Theta_0 > 0}.$$

Juhul kui Θ_0 on tegelik mediaan (nullhüpoteesi kehtides) siis peaks teststatistiku jaotus olema binoomjaotus, $B(X, \Theta_0) \stackrel{H_0}{\sim} B(n; 0,5)$. Alternatiivina võime kasutada teststatistikuna suurust

$$S(X, \Theta_0) = \sum_{i=1}^n \text{sgn}(X_i - \Theta_0).$$

Ka viimase jaotuse saame leida binoomjaotust kasutades, kui paneme tähele, et $S(X, \Theta_0) = 2B(X, \Theta_0) - n$.

Näide 1.1 Vaatame valimit suurusega $n = 10$. Statistiku $S(X, \Theta_0)$ jaotus nullhüpoteesi $\text{med}(X) = \Theta_0$ kehtides on leitav kasutades binoomjaotust $\frac{S(X, \Theta_0) + n}{2} \sim B(n; 0,5)$ ehk $P(S(X, \Theta_0) = s) = \binom{s/2 + 5}{10} 0,5^{5+s/2} 0,5^{5-s/2}$. Vaata statistiku jaotust tabelist 1.1.

Tabel 1.1: Märgitesti statistiku jaotus nullhüpoteesi kehtides, $n = 10$

b	s	$P(B(X, \theta_0) = b) = P(S(X, \theta_0) = s)$
0	-10	0,001
1	-8	0,010
2	-6	0,044
3	-4	0,117
4	-2	0,205
5	0	0,246
6	2	0,205
7	4	0,117
8	6	0,044
9	8	0,010
10	10	0,001

Näeme, et tõenäosusega 0,89 peaks statistiku väärtused jääma vahemikku -4 kuni 4 , tõenäosusega 0,978 aga vahemikku -6 kuni 6 . Kui statistiku $S(X, \Theta_0)$ väärtus peaks olema -6 väiksem või suurem 6 -st, saame olulisuse nivool $0,05$ väita, et nullhüpotees ei pea paika.

Näide 1.2 *Soovime kontrollida järgmiseid hüpoteese:*

$$H_0 : \text{med}(X) = 4$$

$$H_1 : \text{med}(X) \neq 4.$$

Olgu vaatlusandmed järgmised: 3,38; 3,95; 4,15; 4,46; 4,62; 5,19; 5,44; 5,81; 5,82; 6,56. Neist kaks on väiksemad kui 4, ülejäänud on suuremad, seega

$$S(X, 4) = -1 - 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 = 6.$$

Testi olulisustõenäosuse leidmiseks arvutame, kui tõenäoliselt võiks nullhüpoteesi kehtides näha sedavõrd ekstreemset (või veel ekstreemsemat) teststatistiku väärtust (kasutame näites 1.1 toodud tabelit 1.1):

$$\begin{aligned} p\text{-value} = & \overset{0,001}{P(S = 10)} + \overset{0,010}{P(S = 8)} + \overset{0,044}{P(S = 6)} + \\ & \overset{0,044}{P(S = -6)} + \overset{0,010}{P(S = -8)} + \overset{0,001}{P(S = -10)} = 0,11 \end{aligned}$$

Seega olulisuse nivool 0,05 peaksime jääma nullhüpoteesi juurde (mediaan võib küll olla 4).

Paneme tähele, et üsna analoogsel viisil saab kontrollida hüpoteese mistahes teise populatsiooni kvantiili kohta. Näiteks soovides testida (kasutades statistikut $B(X, \Theta_0)$), kas alumine kvartiil (0,25-kvantiil) võiks olla 4, peame vaid olulisustõenäosuse arvutamisel kasutama binoomjaotust $B(n; 0,75)$ (sest 0,25-kvantiilist suurema väärtuse saamise tõenäosus on 0,75).

1.4 Usaldusintervall mediaanile

Järgneval vaatame, kuidas saab leida mitteparameetrilist usaldusintervalli mediaanile. Usaldusintervalli konstrueerimiseks saab kasutada märgitesti — $1 - \alpha$ -usaldusvahemikku moodustavad kõik need Θ_0 väärtused, mille puhul märgitest olulisuse nivool α otsustab nullhüpoteesi kasuks (kui $\Theta_0 = \Theta$ ehk kui nullhüpotees kehtib, siis ei tohi testi olulisustõenäosus olla väiksem α -st suurema tõenäosusega kui α , seega jääb tegelik mediaan Θ usaldusintervalli vähemalt tõenäosusega $1 - \alpha$).

Näide 1.3 *Vaatame näites 1.2 toodud vaatluseid. Tabelis 1.1 on toodud statistiku $S(X, \Theta_0)$ jaotus (eeldusel, et nullhüpotees kehtib). Näeme, et olulisuse nivool 0,022 jääksime nullhüpoteesi juurde siis, kui $-6 \leq S(X, \Theta_0) \leq 6$.*

Leiame, milliste Θ_0 väärtuste korral me saame märgitesti statistiku väärtuseid piirides $[-6..6]$. See pole eriti keeruline ülesanne, sest märgitesti statistiku väärtus võib muutuda vaid valimi vaatluste poolt määratud punktides, vt ka tabelit 1.2. Selliste Θ_0 väärtuste vahemik, $(3,95..5,82)$ olekski 0,978-usaldusintervall mediaanile. Kuna $P(-4 \leq S(X, \Theta_0) \leq 4) = 0,89$, siis 0,89-usaldusintervalliks on vahemik $(4,15..5,81)$. Märkame, et täpselt 0,9-või 0,95-usaldusintervalli leida pole võimalik.

Tabel 1.2: Usaldusintervall mediaanile, märgitesti abil

Θ_0	$S(X, \Theta_0)$
$-\infty..3,38$	10
$3,38..3,95$	8
$3,95..4,15$	6
$4,15..4,46$	4
$4,46..4,62$	2
$4,62..5,19$	0
$5,19..5,44$	-2
$5,44..5,81$	-4
$5,81..5,82$	-6
$5,82..6,56$	-8
$6,56..\infty$	-10

1.5 Mediaani mitteparameetriline hindamine

Mediaani oskame muidugi juba ammu valimi põhjal hinnata. Siin tutvustame lihtsalt loogikat, mida saab hiljem üldistada ka teiste situatsioonide jaoks.

Meil on statistik, antud juhul $S(X, \Theta_0)$, mille abil saame testida hüpoteese mediaani kohta. Millise statistiku $S(X, \Theta_0)$ väärtuse puhul jääme kõige kindlamini nullhüpoteesi juurde? Selliseks statistiku väärtuseks oleks $S(X, \Theta_0) = 0$, mille puhul hüpoteesi $H_0 : \Theta = \Theta_0$ olulisustõenäosuseks tuleks 1. Seega kõige paremini sobivaks hinnanguks Θ väärtusele on arv $\tilde{\Theta}$, mille puhul $S(X, \Theta) = 0$. Teine argument, millega jõutakse sama tulemuseni, kõlab järgmiselt. Kui Θ on tegelik mediaan, siis $ES(X, \Theta) = 0$ (sest binoomjaotus on sümmeetriline) ja momentide meetodit kasutades võiksime valida mediaani hinnanguks $\tilde{\Theta}$ sellise väärtuse, mille puhul $S(X, \tilde{\Theta}) = 0$ ehk $\tilde{\Theta} \in (4,62 \dots 5,19)$. Millist konkreetset väärtust antud vahemikust valida punkthinnanguks, on juba kokkuleppe küsimus. Üheks enam-vähem loogiliseks valikuvõimaluseks oleks valida lõigu keskpunkt $(4,62 + 5,19)/2 = 4,905$, milline olekski üks võimalik mitteparameetriline hinnang mediaanile.

Peatükk 2

Wilcoxon astakmärgitest

2.1 Motivatsioon ja teststatistik

Wilcoxon astakmärgitesti (*Wilcoxon Signed-Rank Test*) kasutatakse kahe sõltuva valimi võrdlemiseks. Oletame näiteks, et soovime võrrelda, millise väetise abil saame parema saagi. Iga katsepõld jagatakse kaheks pooleks, millest ühte poolt väetatakse ühe väetisega, teist poolt aga teise väetisega. Sügisel vaadatakse kuidas saagiga lood on. Näiteks võis vaatluse all olla kolm põldu ja olgu saadud saagid järgmised:

	I pool	II pool	erinevus
1. põld	76	78	2
2. põld	82	91	9
3. põld	80	86	6

Kahel põllul oli parem uue väetisega väetatud pool, ühel põllul andis parema tulemuse vana väetis. Kas see on ükskõik, millisel põllul oli vana väetis parem? Vaatame kahte hüpoteetilist olukorda. Ühel juhul on uus väetis paaril põllul mäekõrguselt parem vanast (6 ja 9 ühikut) ja ühel põllul oli vana väetisega väetatud pool napilt parem (2 ühikut). Teisel juhul annab uue väetisega väetatud pool kahel põllul napilt paremat saaki (2 ja 6 ühikut), üks vana väetisega väetatud põllupool on aga märkimisväärselt parem kui uue väetisega väetatud põllupool (9 ühikut). Kas mõlemal kirjeldatud juhul on andmed uue väetise kasulikkusest sama tugevad? Või arvame, et esimesena kirjeldatud juhul usume meelsamini uue väetise paremusse kui teisena kirjeldatud juhu puhul?

Loomulikult võime taoliste andmete puhul kasutada märgitesti. Leiame iga põllu jaoks saagikuste erinevuse (uue väetisega poole saak - vana väeti-

sega poole saak), tähistame saadud saagikuste erinevuse i . põllul sümboliga X_i . Võime kontrollida märgitesti abil, kasutades teststatistikut $S(X,0) = \sum_{i=1}^n \text{sgn}(X_i)$, kas erinevuste mediaan võiks olla null.

Aga võime ka jälgida oma sisetunnet ja otsida midagi märgitestist paremat. Sellist, mis arvestaks mingil moel ka saagikuste erinevuse suurust. Wilcoxon'i astakmärgitesti statistik võtabki arvesse lisaks erinevuse märgile ka erinevuse (absoluutväärtuse) suurust (astakut):

$$W = \sum_{i=1}^n \text{sgn}(X_i) \text{rank}(|X_i|).$$

2.2 Statistiku W jaotus nullhüpoteesi kehtides

Kui uus ja vana väetis on võrdväärselt head, siis tulenevad erinevused saakides lihtsalt põllupoolte erinevusest. Kumba põllupoolt väetatakse uue väetisega, on heas katses valitud juhuslikult (otsustatakse mündiviske abil vms). Kui vaatame saagikuste erinevusi — saak uue väetisega miinus saak vana väetisega — siis nullhüpoteesi kehtides peaksid märgid erinevuste ees olema täiesti juhuslikud. Antud näites on vaadeldud kolme põldu, seega on võimalikke märgikombinatsioone erinevuste ees $2^3 = 8$, ning järgnevad katsetulemused (valimid) peaksid kõik esinema võrdse tõenäosusega:

erinevused (x_i)			$\text{sgn}(x_i)\text{rank}(x_i)$			$W(x_i)$	$P(W = W(x_i))$
-2	-6	-9	-1	-2	-3	$W = -6$	1/8
+2	-6	-9	+1	-2	-3	$W = -4$	1/8
-2	+6	-9	-1	+2	-3	$W = -2$	1/8
+2	+6	-9	+1	+2	-3	$W = 0$	1/8
-2	-6	+9	-1	-2	+3	$W = 0$	1/8
+2	-6	+9	+1	-2	+3	$W = +2$	1/8
-2	+6	+9	-1	+2	+3	$W = +4$	1/8
+2	+6	+9	+1	+2	+3	$W = +6$	1/8

Ülaltoodud tabelit kasutades võime välja kirjutada Wilcoxon'i astakmärgitesti statistiku W jaotuse juhul, kui $n = 3$, vaata tabelit 2.1.

Oletame nüüd, et meie katses tulid erinevused 6 ja 9 uue väetise kasuks ning erinevus 2 vana väetise kasuks. Seega $W = 4$. Leiame olulisustõenäosuse (p-value) — milline on tõenäosus näha sedavõrd ekstremaalset või veel ekstremaalsemat statistiku väärtust nullhüpoteesi kehtides? Ühepoolse hüpoteesi korral, kui uus väetis saab olla vaid kas samahea või parem kui vana väetis,

Tabel 2.1: Statistiku W jaotus nullhüpoteesi kehtides, $n = 3$

w	-6	-4	-2	0	2	4	6
P(W=w)	1/8	1/8	1/8	2/8	1/8	1/8	1/8

oleksid sama ekstreemse ($W = 4$) või veel ekstreemsema ($W = 6$) statistiku väärtuse saamise tõenäosus nullhüpoteesi kehtides $1/8 + 1/8 = 0,25$, seega olulisustõenäosus on ühepoolse hüpoteesi korral $0,25$. Kahepoolse hüpoteesi korral (kui uus väetis võib põhimõtteliselt olla ka kehvem kui vana) on sama ekstreemse või veel ekstreemsemate statistiku väärtuste hulk $\{-6; -4; 4; 6\}$ suurem ja testi olulisustõenäosuseks tuleb $0,25 * 2 = 0,5$. Seega antud juhul ei suuda me alternatiivset hüpoteesi tõestatuks lugeda mistahes mõistlikku olulisuse nivood kasutades.

Võime ka tähele panna, et kolme vaatluse korral poleks ühegi katsetulemuse korral võimalik tõestada uue väetise paremust vanast olulisuse nivool $0,05$!

Seega üldjuhul on võimalik Wilcoxon astakmärgitesti olulisustõenäosust leida näiteks nii: vaatame läbi 2^n erinevat märgikombinatsiooni astakute ees, leiame iga kombinatsiooni jaoks statistiku W väärtuse ning kirjutame välja W jaotuse juhul, kui nullhüpotees kehtib. Saadud jaotust kasutades on juba lihtne leida testi olulisustõenäosust. Tasub mees pidada, et iga valimi suuruse n korral tuleb statistiku W jaotus erinev!

Samas pole olulisustõenäosuse leidmiseks tarvis tervet jaotust välja kirjutada. Piisab, kui loeme kokku ainult nähtud W väärtuse ja temast veel ekstreemsemate väärtuste saamise tõenäosused nullhüpoteesi kehtides.

Vaatame järgmist näidet

Näide 2.1 Ravimifirma soovis uurida, kas tema poolt välja töötatud spetsiaalset haavaplaastrit kasutades paranevad haavad kiiremini kui traditsioonilist ravimeetodit (haavade kokkuõmblemist) kasutades. Uuringusse värvati 10 katsealust (rotti) ja igale rotile tehti kaks samapikka sisselõiget. Üks haavadeist õmmeldi kokku (kasutati traditsioonilist raviviisi) ja teine plaasterdati kinni. Lasti siis haavadel paar päeva paraneda. Haava paranemise headust mõõdeti paari päeva pärast järgmiselt: vaadati, kui tugevat jõudu läheb vaja haava uuesti lahti rebimiseks (seega suuremad numbrid näitavad paremini paranenud haavu). Katsetulemused on toodud tabelis 2.2.

Leiame kogutud andmete põhjal Wilcoxon astakmärgitesti statistiku vää-

Tabel 2.2: Kliiniline katse parima haavaravi leidmiseks

rott	plaaster	õmblus	plaaster-õmblus	$\text{sgn}(x_i)\text{rank}(x_i)$
1	659	452	207	6
2	984	587	397	9
3	397	460	-63	-1
4	574	787	-213	-7
5	447	351	96	4
6	479	277	202	5
7	676	234	442	10
8	761	516	245	8
9	647	577	70	3
10	577	513	64	2

tuse:

$$W = 6 + 9 - 1 - 7 + 4 + 5 + 10 + 8 + 3 + 2 (= 39).$$

Antud juhul on erinevaid võimalusi astakute ette märke panna $2^{10} = 1024$, kõiki neid läbi vaadata on raske. Aga vaatame, mitu märgikombinatsiooni on sellised, mille puhul tuleks teststatistiku väärtus 39 või suurem. Astakud, mille ette miinusmärgi pannes saame sama ekstreemse või veel ekstreemsema statistiku väärtuse, on ära toodud alljärgnevas tabelis:

miinusmärkide arv	astakud	juhte kokku
0	$\emptyset$	1
1	1, 2, 3, 4, 5, 6, 7, 8	8
2	(1; 2), (1; 3), (1; 4), (1; 5), (1; 6), (1; 7), (2; 3), (2; 4), (2; 5), (2; 6), (3; 4), (3; 5)	12
3	(1; 2; 3), (1; 2; 4), (1; 2; 5), (1; 3; 4)	4

Kokku on seega neist 1024-st juhust selliseid, mille puhul W väärtus tuleks sama ekstreemne või veelgi ekstreemsem kui vaadeldud statistiku väärtus $1 + 8 + 12 + 4 = 25$. Kui me teaksime ette (juba enne andmete kogumist/nägemist), et plaaster saab õmblusest vaid parem olla (mitte halvem), võiksime kontrollida ühepoolset hüpoteesi. Sellisel juhul oleks sedavõrd ekstreemse või veel ekstreemsema statistiku väärtuse nägemise tõenäosus nullhüpoteesi kehtides (p -value) $25/1024 = 0,0244 \dots$. Kui me ei saa kindlalt öelda, et "plastriga haavad ei saa paraneda kehvemini kui õmblust kasutades", peaksime kasutama kahepoolset hüpoteesi. Sel juhul saaksime olulisustõenäosuseks $0,0244 * 2 = 0,0488$. Antud näite korral saame seega tõestatuks lugeda (olulisuse nivool 0,05), et plaaster parandab haavu paremini.

Juhul kui vaatluste arv läheb suureks, võib täpse olulisustõenäosuse leidmine isegi arvuti abil muutuda keeruliseks (näiteks täiesti mõistliku valimi suuruse $n = 100$ korral peaksime läbi vaatama $\approx 10^{30}$ erinevat märgikombinatsiooni...). Isegi siis, kui peaksime olulisustõenäosuse leidmiseks läbi vaatama vaid umbes 5% neist, oleks seda tänapäeva arvutite jaoks liiga palju. Seetõttu kasutatakse vähegi suuremate valimimahtude korral ligikaudseid meetodeid olulisustõenäosuse leidmiseks. Üheks võimaluseks oleks Monte-Carlo kursusest tuttavate meetodite kasutamine. Teiseks võimaluseks oleks statistiku W asümptootilise jaotuse kasutamine.

2.2.1 Monte-Carlo meetod olulisustõenäosuse arvutamiseks

Meetodi idee on lihtne. Iga vaadeldud põllu puhul viskame juhuslikult kulli ja kirja ning otsustame takkajärgi, kumb põllupool peaks kandma silti “uus väetis”. Saagikuste andmed jätame samaks — seega otsustab mündivise iga erinevuse ees kõigest tema märgi. Saagikuse ja väetise tüübi vahel seost olla ei tohi — oleme ju väetise tüüpi näitavad “sildid” juhuslikul segamini ajanud. Arvutame Wilcoxon astakmargitesti statistiku väärtuse sellise segaminiaetud andmestiku pealt. Kordame seda protseduuri palju kordi, näiteks miljon korda. Saame miljon statistiku W väärtust olukorras, kus nullhüpotees garanteeritult kehtib (tänu sellele, et vahetasime põllupooltel silte juhuslikult). Vaatame, kui sageli tuli meie poolt genereeritud andmestikes statistiku väärtus sama erakordne või veelgi erakordsem kui meie valimis. Saadud suhteline sagedus (sama veidrate või veelgi veidramate statistikute arv/ genereeritud valimite arv) olekski olulisustõenäosuse hinnanguks. Olulisustõenäosuse täpsemaks hindamiseks peaksime lihtsalt genereerima rohkem valimeid.

2.3 Asümptootiline jaotus

Astakmargitesti statistik on kirja pandav kui juhuslike suuruste summa, $W = \sum_{i=1}^n Z_i$, kus $Z_i := \text{sgn}(X_i)\text{rank}(|X_i|)$. Juhul kui igal uuritaval objektil on valitud juhuslikult millises järjekorras me töötluseid rakendame (juhuslikult valime kumba haavadest saab plaastri; valime juhuslikult põllupoolle kuhu külvame sorti A jne) saame liidetavaid lugeda sõltumatuteks ning sõltumatute juhuslike suuruste summa on üldjuhul asümptootiliselt normaaljaotusega.

Leiame esmalt statistiku W keskväärtuse ja dispersiooni H_0 kehtides. Eldame, et nullhüpotees kehtib. Sellisel juhul tuleneb statistiku juhuslikkus puhtalt töötluste randomiseerimisest, seega suurused $\text{rank}(|X_i|)$ on käsitle-

tavad kui fikseeritud konstandid ja juhuslikud on ainult suurused $\text{sgn}(X_i)$:

$$\begin{aligned} P(\text{sgn}(X_i) = -1) &= 0,5 \\ P(\text{sgn}(X_i) = 1) &= 0,5 \end{aligned}$$

Seega (nullhüpoteesi kehtides):

$$\begin{aligned} EW &= E \sum_{i=1}^n Z_i \\ &= \sum_{i=1}^n \text{rank}(|X_i|) E(\text{sgn}(X_i)) \\ &= \sum_{i=1}^n \text{rank}(|X_i|) \cdot 0 = 0 \end{aligned}$$

ja

$$\begin{aligned} DW &= D \sum_{i=1}^n Z_i \\ &= \sum_{i=1}^n \text{rank}^2(|X_i|) D(\text{sgn}(X_i)) \\ &= \sum_{i=1}^n \text{rank}^2(|X_i|) \\ &= 1^2 + 2^2 + 3^2 + \dots n^2 \\ &= n(n+1)(2n+1)/6 \end{aligned}$$

Võrduse $\sum_{i=1}^n i^2 = n(n+1)(2n+1)/6$ kehtivuses veendumiseks võib kasutada näiteks induktsiooni. Veendu, et antud valem kehtib, kui $n = 1$. Siis näita, et kui valem kehtib n korral, siis kehtib ta ka $n + 1$ korral:

$$\begin{aligned} \sum_{i=1}^{n+1} i^2 &= \sum_{i=1}^n i^2 + (n+1)^2 \\ &= n(n+1)(2n+1)/6 + (n+1)^2 \\ &= (n+1)(2n^2 + n)/6 + (n+1)^2 \\ &= (n+1)(2n^2 + n + 6n + 6)/6 \\ &= (n+1)(2n^2 + 7n + 6)/6 \\ &= (n+1)(n+2)(2(n+1) + 1)/6. \end{aligned}$$

Statistiku W liidetavad on küll sõltumatud, kuid pole sama jaotusega. Siiski saab tsentraalse piirteoreemi üldistuse (Lindeberg-Felleri teoreem) põhjal väita, et standardiseeritud statistiku

$$\frac{W - EW}{\sqrt{DW}}$$

jaotus läheneb nullhüpoteesi kehtides standardsele normaaljaotusele:

$$P\left(\frac{W}{\sqrt{\frac{n(n+1)(2n+1)}{6}}} < a\right) \rightarrow \Phi(a), \quad \text{kui } n \rightarrow \infty.$$

Näide 2.2 Proovitakse, kas päikesekreem ikka kaitseb nahka päikese eest. Katsealustel määratakse lapike nahka kreemiga kokku. Saadetakse nad siis randa. Hiljem mõõdetakse naha punetust nii kreemitatud kohast kui ka selle kõrvalt. Naha punetuse erinevused (kaitsmata koht-kreemitatud koht) on järgmised:

-1; -3; 4; -6; 7; 8; -9; -10; 12; 14; 15; -17; 20; 25; 29; 30; -32; 40; 78; 102.

Wilcoxon astakmärgitesti statistiku väärtuseks saame:

$$\begin{aligned} W &= -1 - 2 + 3 - 4 + 5 + 6 - 7 - 8 + 9 + 10 + \\ &\quad + 11 - 12 + 13 + 14 + 15 + 16 - 17 + 18 + 19 + 20 = 108 \end{aligned}$$

Nullhüpoteesi kehtides ootaksime antud valimi suuruse juures statistiku kesk-väärtuseks ja dispersiooniks $EW = 0$ ja $DW = n(n+1)(2n+1)/6 = 20 * 21 * 41/6 = 2870$. Seega peaks nullhüpoteesi kehtides juhuslik suurus $(W - EW)/\sqrt{DW} = W/53,57$ olema ligikaudu standardse normaaljaotusega juhuslik suurus (nullhüpoteesi kehtides peaks vastav suurus ootuspäraselt jääma vahemikku -1,96...1,96). Antud valimi korral aga $W/53,57 = 108/53,57 = 2,016$, seega midagi tundub olevat viltu... nullhüpoteesi eeldusega. Olulisustõenäosuse saame, kui leiame, kui tõenäoliselt oleksime näinud sedavõrd ekstreemset või veel ekstreemsemat statistiku väärtust nullhüpoteesi kehtides. Tõenäosus näha veel suuremat statistiku väärtust H_0 kehtides oleks $1 - \Phi(2,016) = 0,0219$ (olulisustõenäosus ühepoolse hüpoteesi korral). Kahepoolse hüpoteesi korral peaksime muidugi siia juurde liitma ka erakordselt väikese W saamise tõenäosuse (nullhüpoteesi kehtides). Normaaljaotuse sümmeetriat arvestades saame tulemuseks $(1 - \Phi(2,016)) * 2 = 0,0438$. Seega on olulisustõenäosus väike ja me võime nullhüpoteesi (olulisuse nivool 0,05) kummutatuks lugeda. Võrdlusena: täpne olulisustõenäosus oleks 0,044054.

2.3.1 Pidevuse parandus

Statistikas tuleb sageli ette olukordi, kus me lähendame diskreetset jaotust pideva jaotusega. Näiteks võime lähendada binoomjaotust normaaljaotusega või Wilcoxon'i astakmargitesti statistiku jaotust normaaljaotusega. Kui meie valimid on väga suured, siis võime sellist lähendamist teha üsna sirgjooneliselt ja probleemideta. Keskmise suurusega valimite korral saame aga sageli veidi täpsemaid tulemusi, kui enne arvutuste alustamist veidi mõtleme.

Vaatame näiteks juhuslikku suurust $X \sim B(10; 0,4)$. Antud binoomjaotust saab lähendada normaaljaotusega juhusliku suurusega $Y \sim N(4; \sigma^2 = 2,4)$. Kuidas leida normaaljaotuse lähendit kasutades tõenäosust, et $X = 1$, $P(X = 1) = 0,0403$? On selge, et antud tõenäosus pole ligilähedaseltki võrdne tõenäosusega $P(Y = 1) = 0$. Parema variant antud tõenäosuse leidmiseks oleks vaadata tõenäosust $P(0,5 \leq Y \leq 1,5) = 0,0414$.

Kuidas aga oleks mõistlik leida normaaljaotuse lähendit kasutades suurust $P(X \leq 1) \approx 0,04636$? Me võime mõelda ka nii: $P(X \leq 1) = P(X = 1) + P(X = 0)$. Viimast tõenäosust aga lähendab tõenäosus $P(Y \leq 1,5) = 0,0533$ palju täpsemalt kui tõenäosus $P(Y \leq 1) = 0,026$. Seega soovitatakse kasutada diskreetsete juhuslike suuruste lähendamisel pidevuse korrektsiooni (*continuity correction*): $P(X \leq c) \approx P(Y \leq c + 0,5)$ või $P(X \geq c) \approx P(Y \geq c - 0,5)$.

Vaatame pidevuse parandust eelnenud näite varal. Leidmaks tõenäosust $P(W \geq 108)$ oleks meil seega soovitatav vaadata normaaljaotuse jaotusfunktsiooni kohal $(W - 1)/53,57$ (kuna W muutub meil sammuga 2), mis annaks kahepoolse hüpoteesi puhul olulisustõenäosuseks $(1 - \Phi(1,9974)) * 2 = 0,0458$. Antud näites ei andnud pidevuse paranduse kasutamine täpsemat tulemust — kui ilma pidevuse paranduseta alahindasime olulisustõenäosust, siis pidevuse parandust kasutades saime liiga suure hinnangu olulisustõenäosusele. Lähend on kõigest lähend ja vahel võib pidevuse paranduse kasutamine ka halvemaid tulemusi anda. Enamasti saadakse siiski pidevuse parandust kasutades täpsemaid tulemusi (ja ka antud näite puhul on arvatavasti konservatiivsem hinnang olulisustõenäosusele eelistatum).

2.4 Võrdsed astakud

Pideva tunnuse korral on kahe täpselt võrdse vaatluse saamise tõenäosus 0. Paraku ei taha aga praktika sugugi kuulata teooria sõna. Mõõtmiste ebatäpsusest, vahetulemuste ümmardamisest vms tingitult esineb ikka samasuuri väärtuseid.

Kui mõlema töötluse korral saadakse samasuur uuritava tunnuse väärtus (ei oska öelda, kumb töötlus andis parema tulemuse), ehk töötluste erinevus on 0, siis taolised vaatlused visatakse tüüpiliselt lihtsalt testimise ajaks valimist välja (vähendatakse valimi suurust).

Kui meil on kahel uuritaval indiviidil toimunud (mõõtmistäpsuse piires) samasuur muutus (muutuse absoluutväärtust arvestades) ei oska me ühele muutusele teisest kõrgemat kohta (astakut) omistada. Lahendus on muidugi üsna lihtne — anname kõigile samasuurtele muutustele ühe ja sama astaku väärtuse, milleks oleks nende astakute keskmine. Näiteks kui me ei suuda otsustada, kumb vaatlustest pretendeerib 4. ja kumb 5. kohale, siis anname mõlemale vaatlusele astakuks $(4 + 5)/2 = 4,5$. Sellisel juhul on statistik defineeritud valemiga $W^* = \sum_{i=1}^n \text{sgn}(X_i) \text{midrank}(|X_i|)$.

Muutes taolisel viisil astakuid muutuvad muidugi ka väärtused, milliseid statistik omandada võib. Seega muutub ka Wilcoxon statistiku jaotus nullhüpoteesi kehtides. Kui vaatame ise paberil läbi statistiku kõik võimalikud väärtused, siis arvatavasti ei teki sellest suuremat probleemi — statistiku jaotustabel tuleb lihtsalt veidi teistsugune kui siis, kui võrdseid vahesid poleks esinenud.

Soovides kasutada statistiku W^* asümptootilist jaotust, peame samuti arvestama võrdsete vaatlustega — need muudavad mõnevõrra statistiku asümptootilist jaotust.

On lihtne näha, et keskväärtus ei muutu — nullhüpoteesi kehtides jätkuvalt $EW^* = 0$. Küll aga muutub Wilcoxon astakmärgite testi statistiku dispersioon (erinevate väärtuste arv kahaneb — järelikult kahaneb ka varieeruvus). Dispersioon DW^* avaldub:

$$\begin{aligned} DW^* &= D \sum_{i=1}^n Z_i^* \\ &= \sum_{i=1}^n \text{midrank}^2(|X_i|) D(\text{sgn}(X_i)) \\ &= \sum_{i=1}^n \text{midrank}^2(|X_i|) \end{aligned}$$

Liites ja lahutades eelmisele avaldisele nõ tavalise Wilcoxon astakmärgi-testi statistiku dispersiooni $DW = \sum_{i=1}^n \text{rank}^2(|X_i|)$ saame dispersiooni DW^* esitada järgmisel kujul:

$$\begin{aligned}
 DW^* &= \sum_{i=1}^n \text{midrank}^2(|X_i|) + \sum_{i=1}^n \text{rank}^2(|X_i|) - \sum_{i=1}^n \text{rank}^2(|X_i|) \\
 &= \sum_{i=1}^n \text{rank}^2(|X_i|) - \left(\sum_{i=1}^n \text{rank}^2(|X_i|) - \sum_{i=1}^n \text{midrank}^2(|X_i|) \right) \\
 &= \sum_{i=1}^n \text{rank}^2(|X_i|) - \text{muutus} \\
 &= DW - \text{muutus}.
 \end{aligned}$$

Edasi leiame, kui suur on tegur 'muutus', mis näitab astakute ruutude summas toimunud muutust. Oletame esmalt, et meil esineb võrdse suurusega vaatluseid vaid ühes kohas — vaatlused $(s+1)$ kuni $(s+t)$ olgu kõik sama-suured. Nende vaatluste keskmiseks astakuks oleks $s + (t+1)/2$ ja taoliste astakute ruutude summa on

$$t \left(s + \frac{t+1}{2} \right)^2 = t \left(s^2 + s(t+1) + \frac{(t+1)^2}{4} \right).$$

Juhul, kui võrdse suurusega vaatluseid poleks, oleks nendesamade astakute ruutude summaks olnud

$$\sum_{i=1}^t (s+i)^2 = ts^2 + 2st(t+1)/2 + t(t+1)(2t+1)/6.$$

Leiame, kui palju muutus astakute ruutude summa:

$$\begin{aligned}
 \text{muutus} &= ts^2 + 2st(t+1)/2 + t(t+1)(2t+1)/6 - \\
 &\quad t \left(s^2 + s(t+1) + \frac{(t+1)^2}{4} \right) \\
 &= t(t+1)(2t+1)/6 - t(t+1)^2/4 \\
 &= t(t+1)[(2t+1) * 2 - (t+1) * 3]/12 \\
 &= t(t+1)(t-1)/12 \\
 &= t(t^2-1)/12.
 \end{aligned}$$

See on astakute ruutude summa erinevus kui korduvaid väärtuseid on variatsioonreas ühes kohas t tükki. Olgu meil valimis N' erinevat väärtust,

kusjuures j . unikaalsed väärtust esinegu t_j korda. Astakute ruutude summa muutus on sel korral

$$\sum_{j=1}^{N'} t_j(t_j^2 - 1)/12.$$

Ja Wilcoxon astakmärgistatistiku W^* dispersioon (nullhüpoteesi kehtides) on leitav järgmise valemi abil:

$$DW^* = \frac{n(n+1)(2n+1)}{6} - \sum_{j=1}^{N'} t_j(t_j^2 - 1)/12.$$

Kui võrdseid vaatluseid on palju, võib dispersioon märkimisväärselt väheneda. NB! Mitte kõik programmid ei pruugi võrdse suurusega vaatluseid kohates kõige mõistlikumalt käituda. Kui võrdseid astakuid on palju, siis võib Wilcoxon astakmärgitesti statistiku jaotus koonduda normaaljaotuseks aeglaselt (vajame tõesti suurt valimit, enne kui võime kasutada lähendit!)

2.5 Erinevuse kirjeldamine

Mis saab siis, kui me nullhüpoteesi ümber lükkame? On ju tore, kui saame öelda: “uus ravi on parem vanast” või “uus väetis on parem kui vana”, aga alati tekib küsimus: Kui palju parem? Kui tänu uuele väetisele saame hektarilt täiendavalt 1 grammi jagu rohkem saaki, siis ei tasu uue väetise eest kallimat hinda maksta. Aga kui tänu uuele ja kallimale väetisele saaksime tonni jagu rohkem saaki, tasuks muutus meile ära. Mõlema kirjeldatud juhu puhul võib Wilcoxon astakmärgitist võtta vastu alternatiivse hüpoteesi (piisavalt suure valimi korral).

Üks võimalus töötlustevahelise erinevuse suuruse kirjeldamiseks oleks järgmine. Me võime arutleda, kui palju me peaksime ühe töötlusega saadud tulemustest maha lahutama, et saada teise töötlusega võrdväärseid (sarnaseid) tulemusi.

Täpsemalt, otsitakse sellist Δ väärtust, et Wilcoxon teststatistik kinnitaks peale Δ juurdeliitmist uuele töötlusele võimalikult hästi nullhüpoteesi ehk

$$W(Y - X + \Delta) = 0.$$

Antud võrrandil on sageli rohkem kui üks lahend, millisel juhul on sobilik raporteerida lahenditeks sobivate väärtuste lõigu keskpunkti.

Näide 2.3 *Vaatlustulemused (töötluste erinevused) on järgmised: $-2; 3; 6; 10; 20$. Wilcoxon teststatistiku väärtuseks on $W = -1 + 2 + 3 + 4 + 5 =$*

13. Selleks, et saaksime tulemuseks $W = 0$, peaksime kõigist erinevustest lahutama 6,5. Sellisel juhul oleksid meie vaatlusteks $-8,5$; $-3,5$; $-0,5$; $3,5$; $13,5$ ja statistiku väärtus $W = -4 - 2,5 - 1 + 2,5 + 5 = 0$. Seega $\Delta = 6,5$.

Märkus: Otsitav väärtus, Δ , ei ole vahede (töötluste erinevuste) mediaan!

Tehniline märkus: Kuidas leida sobivat Δ väärtust? Üks võimalus oleks järgmine: leia kõikvõimalikud vahede keskmised (neid suurusid nimetatakse ka Walshi keskmisteks):

$$\frac{(-2) + (-2)}{2}, \frac{(-2) + 3}{2}, \dots, \frac{(-2) + 20}{2}, \frac{3 + 3}{2}, \dots, \frac{3 + 20}{2}, \dots, \frac{20 + 20}{2}$$

ning leia siis nende $5 \cdot (5 + 1)/2$ keskmise mediaan. Viimane ongi otsitavaks Δ väärtuseks! Kas oskad põhjendada, miks?

Saadud Δ väärtuse interpreteerimine võib ja ei pruugi raskuseid valmistada. Kui töötlused erineksid selle poolest, et üks töötlus lisaks uuritava tunnuse väärtusele alati Δ ühikut juurde (nihutaks uuritava tunnuse väärtuseid: $Y_i = X_i + \Delta$) siis on meie saadud Δ väärtus hinnanguks sellele nihkele. Enamasti pole muidugi töötluse mõju kõigile ühesugune (väetamine aitab mõnda põldu enam kui teist; tänu ravile paraneb mõni inimene kibekiiresti ja teine (haigusest rohkem kurnatud) jääb veel mitmeks päevaks voodisse jne. Seega on igal inimesel/põllul oma isiklik töötluse mõju Δ_i . Juhul, kui juhuslike suuruste Δ_i jaotus oleks sümmeetriline, võiksime enda poolt leitud Δ väärtust käsitleda kui hinnangut juhuslike suuruste Δ_i 'de mediaanile. Aga kas töötluste erinevuste jaotus on ikka sümmeetriline? Märkusena olgu mainitud: kui töötlused on eristamatud, siis on erinevuste jaotus $X - Y$ sümmeetriline.

Heal lapsel on ikka mitu nime. Meie poolt vaadeldud Δ 't võib kirjanduses leida nii pseudomediaani (pseudomedian) kui ka Hodges-Lehmann'i hinnangu nime all.

Kui märgitesti statistikut kasutades saime leida usaldusintervalli mediaanile, nii saab Wilcoxon'i astakmärgitesti kasutades leida usaldusintervalli pseudomediaanile. Tuleb lihtsalt leida nihete võimalike väärtuste hulk, mille puhul Wilcoxon'i astakmärgitest jääb (antud olulisuse nivool) veel nullhüpoteesi juurde.

Peatükk 3

Wilcoxon astaksummatest

3.1 Teststatistik

Wilcoxon astaksummatest (*Wilcoxon Rank-Sum Test*) on sõltumatuid valimeid kasutav mitteparameetriline test kahe populatsiooni võrdlemiseks.

Alustuseks vaatame järgmist näidet. Peame ravima viit gripihaiget. Soovime teada saada, kas erilise hiina tee abil saaksid haiged kiiremini terveks. Jagame viis haiget juhuslikult kahte rühma — ravi saavaks ja kontrollrühmaks (nemad saavad tavalist teed). Ravi saavasse rühma valime juhuslikult 3 patsienti, kontrollrühma jääb 2 patsienti. Jälgime patsiente ja märgime üles, kuna nad terveks saavad. Esimesena terveks saanud patsiendi tähistame numbriga 1, teisena terveks saanud haige tähistame numbriga 2 jne.

Haigetele omistatud tervenemisaja astakud võivad jaguneda kahte gruppi 10-l erineval viisil ($C_5^3 = C_5^2 = 10$):

Ravi	1, 2, 3	1, 2, 4	1, 2, 5	1, 3, 4	1, 3, 5
Kontroll	4, 5	3, 5	3, 4	2, 5	2, 4

Ravi	1, 4, 5	2, 3, 4	2, 3, 5	2, 4, 5	3, 4, 5
Kontroll	2, 3	1, 5	1, 4	1, 3	1, 2

Juhul, kui ravil puuduks igasugune mõju, oleks kõigi võimalike järjestuste esinemistõenäosus ühesugune — kõik vaadeldud astakute jagunemised gruppidesse leiavad aset tõenäosusega $1/10$. Kui aga ravi aitaks patsiente, siis kipuksid ravigrupis astakud olema väiksemad kui kontrollgrupis (ravi-grupis saavad inimesed varem terveks). Siit tuleneb ka idee teststatistiku valikuks: liidame kõigi ravigruppi sattunud inimeste tervenemisaegade as-

takud. Kui saadav summa tuleb väga väike, siis vihjab see võimalusele, et ravist on kasu.

Formaalselt kirja pandult: olgu ravigrupis uuritava tunnuse väärtuste astakud $s_1, s_2, \dots, s_n$ (ravi saanute rühmas oli kokku n inimest). Siis teststatistikuna kasutame suurust

$$W_s = s_1 + s_2 + \dots + s_n.$$

Milline on statistiku W_s jaotus meie näite korral nullhüpoteesi kehtides? Leiame kõigi 10 variandi jaoks statistiku W_s väärtused ja saamegi kirja panna tema jaotuse nullhüpoteesi kehtides (võttes arvesse, et iga võimaliku kombinatsiooni esinemistõenäosus on $1/10$):

w	6	7	8	9	10	11	12
$P(W_s = w)$	0,1	0,1	0,2	0,2	0,2	0,1	0,1

Enne uuringuandmete kogumist oleks tarvis määrata ka kasutatav olulisuse nivoo. Kuna tüüpilise olulisuse nivoo 0,05 korral me nullhüpoteesi kummutada ei saaks (nii väikse valimi korral), siis valigem olulisuse nivoo $\alpha = 0,1$ (ja kasutame patuga pooleks ühepoolset hüpoteesi). Seega kui saaksime teststatistiku väärtuseks $W_s = 6$, siis me kummutaks nullhüpoteesi ja ütleksime, et ravi aitab. Muul juhul jääksime nullhüpoteesi juurde.

Nüüd oleme teinud kõik otsused, mida on tarvis teha enne katsega alustamist ja võime asuda oma eksperimenti läbi viima ja andmeid koguma. Oletame, et esimese, teise ja neljandana terveks saanud inimesed kuulusid ravitavate gruppi. Sellisel juhul $W_s = 1 + 2 + 4 = 7$, ühepoolse testi p-väärtus ehk olulisustõenäosus tuleb $P(W_s \leq 7 | H_0) = 0,1 + 0,1 = 0,2$ ja seega peame jääma nullhüpoteesi juurde.

Antud näites vaatlesime ravi saanud inimesterühma astakute summat. Sama hästi oleksime muidugi võinud vaadelda kontrollrühma kuuluvate patsientide tervenemisaegade astakute summat W_r . Paneme tähele, et statistikute W_s ja W_r väärtused on seotud lihtsa valemiga. Olgu kontrollrühmas tehtud m vaatlust ja ravirühmas n vaatlust, kokku mõlemas rühmas tehtud N vaatlust. Siis

$$W_s = 0,5N(N + 1) - W_r,$$

sest kõigi vaatluste astakute summa on $1 + 2 + \dots + N = 0,5N(N + 1)$. Tänu sellisele üks-ühesele vastavusele on ükskõik, kumba rühma me loeme kontrollgrupiks. Tasub siiski täheldada, et statistikute W_s ja W_r jaotused nullhüpoteesi kehtides võivad olla erinevad. Statistiku W_s minimaalne võimalik väärtus on $0,5n(n + 1)$, seevastu statistiku W_r minimaalne võimalik

väärtus on $0,5m(m+1)$. See tekitab raskuseid tabelite koostamisel, ja seetõttu on kasutusele võetud ka transformeeritud statistik W_{xy} , mille puhul mainitud erinevus kaob. Statistiku W_{xy} leidmiseks tuleb Wilcoxon teststatistikust maha lahutada tema minimaalne võimalik väärtus:

$$W_{xy} := W_s - 0.5n(n+1).$$

Seega ülaltoodud näite korral, kus $W_s = 7$, oleks $W_{xy} = 7 - 0.5 * 3 * (3 + 1) = 1$.

Teststatistikut W_{xy} tuntakse ka Mann-Whitney teststatistiku nime all ja temal baseeruvat testi kutsutakse Mann-Whitney testiks. Pole raske näha, et Mann-Whitney testi otsused langevad kokku Wilcoxon (astaksumma) testi omadega. Mann-Whitney teststatistikul on ka oma interpretatsioon. Nimelt võime vaadelda kõiki võimalikke paare (X, Y) , kus X on pärit ühest populatsioonist võetud valimist (ravigrupist) ja Y on pärit teisest populatsioonist võetud valimist (kontrollgrupist). Mann-Whitney teststatistik näitab siis, kui paljudes vaadeldud paarides oli ravigruppi kuulunud patsiendi tulemus suurem (tervenemiseks läks kauem aega) kui kontrollgruppi kuulunud patsiendil,

$$W_{xy} = \#_{i,j} \{X_i > Y_j\}.$$

Kasutatavas näites oleksid kõikvõimalikud paarid järgmised:

Ravigrupi Astakud (X)	kontrollgrupi astakud (Y)	paarid	ravigrupi liige tervenes aeglasemalt
1	3	1 3	-
2	5	1 5	-
4		2 3	-
		2 5	-
		4 3	+
		4 5	-

Paaride arv mille puhul ravigruppi kuulunud patsiendil läks tervenemiseks kauem aega oli $1 = W_{xy}$.

Teoreem 3.1 $W_{xy} = \#_{i,j} \{X_i > Y_j\}$.

Tõestus.

Vaatleme ravirühma astakuid: ravirühma väikseima vaatluse astak $S_{(1)}$ näitab, mitu kontrollrühma patsienti on temast väiksemad $+1$, st kui ravirühma väikseim astak oleks 3, siis oleks kontrollrühmas 2 vaatlust, mis on

temast väiksemad. Seega antud vaatluse abil moodustatud paaridest oleksid 2 sellised, kus kontrollrühma tulemus oleks olnud väiksem. Üldjuhul on siis pisemast ravirühma vaatlusest väiksemaid kontrollrühma vaatluseid $S_{(1)} - 1$ tükki.

Suuruselt järgmise ravirühma vaatluse astak on $S_{(2)}$, sellest vaatlusest väiksemaid kontrollrühma vaatluseid on aga $S_{(2)} - 2$ tükki (ses ravirühma üks vaatlustest oli ka sellest vaatlusest väiksem) jne. Seega:

$$\begin{aligned}(S_{(1)} - 1) + (S_{(2)} - 2) + \dots + (S_{(n)} - n) &= \#_{i,j}\{X_i > Y_j\} \\ W_s - 0,5n(n+1) &= \#_{i,j}\{X_i > Y_j\} \\ W_{xy} &= \#_{i,j}\{X_i > Y_j\}.\end{aligned}$$

q.e.d.

Märkus: Lisaks W_{xy} -le võib kasutada ka statistikut W_{yx} -i:

$$W_{yx} := W_r - 0,5m(m+1) = \#_{i,j}\{Y_i > X_j\}.$$

3.2 Keskväärtus ja dispersioon H_0 kehtides

Alustuseks vaatame, milline on ühe vaatluse astakute jaotus ja kahe vaatluse astakute ühisjaotus H_0 kehtides.

Lemma 3.1 *Kui H_0 kehtib — populatsioonid on eristamatud — siis*

1. *i. vaatluse astaku S_i jaotus on*

$$P(S_i = k) = 1/N, k = 1, 2, \dots, N.$$

2. *i. ja j. ($i, j = 1, 2, \dots, N$) vaatluse ühisjaotuseks on*

$$P(S_i = k, S_j = l) = \begin{cases} 1/[N(N-1)] & i \neq j \\ 0 & i = j \end{cases}.$$

Tõestus.

H_0 kehtides on kõik $N!$ vaatluste järjestust võrdvõimalikud. Neist $(N-1)!$ on sellised, kus $S_i = k$. Seega $P(S_i = k) = (N-1)!/N! = 1/N$. Selliseid järjestusi, kus $S_i = k, S_j = l$ on aga $(N-2)!$ tükki, seega

$$\begin{aligned}P(S_i = k, S_j = l) &= (N-2)!/N! \\ &= 1/[N(N-1)].\end{aligned}$$

Q.e.d.

Lemma 3.2 *Kui H_0 kehtib, siis:*

1. $ES_i = (N + 1)/2$
2. $DS_i = (N^2 - 1)/12$
3. $Cov(S_i, S_j) = -(N + 1)/12$, kui $i \neq j$

Tõestus.

$$\begin{aligned} ES_i &= 1/N(1 + 2 + \dots + N) \\ &= 1/N * N(N + 1)/2 \\ &= (N + 1)/2 \end{aligned}$$

$$\begin{aligned} DS_i &= 1/N(1^2 + 2^2 + \dots + N^2) - (ES_i)^2 \\ &= 1/N * N(N + 1)(2N + 1)/6 - (N + 1)^2/4 \\ &= (4N^2 + 6N + 2)/12 - (3N^2 + 6N + 3)/12 \\ &= (N^2 - 1)/12. \end{aligned}$$

$$\begin{aligned} Cov(S_i, S_j) &= ES_i S_j - ES_i ES_j \\ &= \frac{1}{N(N-1)}(1 \cdot 1 + 1 \cdot 2 + \dots + 1 \cdot N + 2 \cdot 1 + \dots + 2 \cdot N + \dots + N \cdot N - \\ &\quad -(1^2 + 2^2 + \dots + N^2)) - (N + 1)^2/4 \\ &= \frac{1}{N(N-1)} \left(\frac{N(N+1)}{2}(1 + 2 + \dots + N) - N(N+1)(2N+1)/6 \right) - (N+1)^2/4 \\ &= \frac{1}{N(N-1)}(N(N+1)N(N+1)/4 - N(N+1)(2N+1)/6 - (N+1)^2/4) \\ &= \frac{1}{12N(N-1)}N(N+1)[3N(N+1) - 2(2N+1)] - (N+1)^2/4 \\ &= \frac{1}{12N(N-1)}N(N+1)(3N+2)(N-1) - (N+1)^2/4 \\ &= (N+1)(3N+2)/12 - (N+1)^2/4 \\ &= (N+1)(3N+2-3N-3)/12 \\ &= -(N+1)/12. \end{aligned}$$

Q.e.d.

Lemma 3.3 *Kui H_0 kehtib, siis*

1. $EW_s = n(N + 1)/2$;
2. $DW_s = nm(N + 1)/12$.

Tõestus

$$\begin{aligned} EW_s &= E(S_1 + S_2 + \dots + S_n) \\ &= n(N + 1)/2 \end{aligned}$$

$$\begin{aligned} DW_s &= D(S_1 + S_2 + \dots + S_n) \\ &= nD(S_i) + (n^2 - n) \cdot \text{Cov}(S_i, S_j) \\ &= n(N^2 - 1)/12 - n(n - 1)(N + 1)/12 \\ &= n(N + 1)[N - 1 - (n - 1)]/12 \\ &= nm(N + 1)/12. \end{aligned}$$

Q.e.d.

Märkus. Mann-Whitney teststatistiku keskväärtus ja dispersioon on H_0 kehtides järgmised:

$$\begin{aligned} (EW_{yx} =) EW_{xy} &= mn/2 \\ (DW_{yx} =) DW_{xy} &= nm(N + 1)/12. \end{aligned}$$

3.3 Asümptootiline jaotus

Suure katse puhul, kus n ja m on mõlemad suured, võib tõenäosuse $P(W_s > w)$ arvutamine kõigi võimaluste kokkulugemise teel osutuda väga töömahukaks. Sestap oleks hea, kui teaksime, millisele piirjaotusele läheneb meid huvitava statistiku W_s jaotus (nullhüpoteesi kehtides) suure valimi korral. Selgub, et selleks jaotuseks on normaaljaotus. Antud väite tõestus pole päris triviaalne, sest astakud pole sõltumatud ja seega on W_s sõltuvate juhuslike suuruste summa (tõestus pole ka lootusetult keeruline aga jääb siiski praegu välja). Teisisõnu öeldes, asümptootiliselt:

$$\frac{W_s - E W_s}{\sqrt{D W_s}} \xrightarrow{\mathcal{D}, H_0} \mathcal{N}(0,1).$$

Ehk, asendades keskvaartuse ja dispersiooni ülaltoodud valemisse, saame:

$$\frac{W_s - n(N+1)/2}{\sqrt{mn(N+1)/12}} \xrightarrow{\mathcal{D}, H_0} \mathcal{N}(0,1).$$

Kontrollides ühepoolset hüpoteesi (teame eelnevalt, et töötlus ei mõju või kui mõjub, siis ainult uuritava tunnuse väärtuseid kahandavalt), saame Wilcoxon testi olulisustõenäosuse leida suure valimi korral järgmise valemi abil:

$$\begin{aligned} P(W_s \leq c | H_0) &= \Phi \left(\frac{c + \text{pidevuse parandus} - E_0 W_s}{\sqrt{D_0 W_s}} \right) \\ &= \Phi \left(\frac{c + 0,5 - 0,5n(N+1)}{\sqrt{mn(N+1)/12}} \right) \end{aligned}$$

kus $\Phi(x)$ on standardse normaaljaotuse jaotusfunktsioon kohal x ning $E_0 W_s = E(W_s | H_0)$ ja $D_0 W_s = D(W_s | H_0)$.

Kahepoolne hüpotees.

Oletame esmalt, et vaadeldavas valimis on astaksummatesti teststatistiku väärtus c väiksem kui nullhüpoteesi puhul oodatav teststatistiku väärtus, $c < E_0 W_s$ ehk $c < 0,5n(N+1)$.

Tõenäosus näha sedavõrd väikest või veel väiksemat teststatistiku väärtust on

$$\Phi \left(\frac{c + 0,5 - 0,5n(N+1)}{\sqrt{mn(N+1)/12}} \right).$$

Kahepoolse hüpoteesi puhul pakuvad aga huvi ka sedavõrd oodatavast suuremad või veel suuremad statistiku väärtused. Kui praegu on valimis nähtud teststatistiku väärtus $E_0W_s - c = 0,5n(N+1) - c$ ühikut väiksem H_0 puhul oodatavast, siis kui tõenäoline on näha samavõrra H_0 puhul oodatavast väärtusest suuremat (või veel suuremat) teststatistiku väärtust? Vastav tõenäosus on leitav valemiga

$$\begin{aligned} P(W_s \geq E_0W_s + (E_0W_s - c) | H_0) &= \\ &= P(W_s \geq n(N+1) - c | H_0) \\ &= 1 - P(W_s < n(N+1) - c | H_0) \\ &= 1 - \Phi\left(\frac{n(N+1) - c - \text{pidevuse korrektsioon} - E_0W_s}{\sqrt{D_0W_s}}\right) \\ &= 1 - \Phi\left(\frac{0,5n(N+1) - c - 0,5}{\sqrt{D_0W_s}}\right), \end{aligned}$$

kust saame (kuna $1 - \Phi(x) = \Phi(-x)$):

$$P(W_s \geq E_0W_s + (E_0W_s - c) | H_0) = \Phi\left(\frac{c + 0,5 - 0,5n(N+1)}{\sqrt{mn(N+1)/12}}\right).$$

Liites mõlemad tõenäosused (sedavõrd väikese või veel väiksema teststatistiku saamise tõenäosuse ja samavõrra oodatavast suurema või veel suurema teststatistiku saamise tõenäosuse) saame olulisustõenäosuseks

$$\text{p-väärtus} = 2 \cdot \Phi\left(\frac{c + 0,5 - 0,5n(N+1)}{\sqrt{mn(N+1)/12}}\right). \quad (3.1)$$

Kui aga nähtud teststatistik peaks olema suurem kui nullhüpoteesi puhul oodatav väärtus E_0W_s , siis jõuame sarnase arutelu tulemusel järgmise p-väärtuseni:

$$\text{p-väärtus} = 2 \cdot \Phi\left(-\frac{c - 0,5 - 0,5n(N+1)}{\sqrt{mn(N+1)/12}}\right). \quad (3.2)$$

Valemite 3.1 ja 3.2 pealt võime kokku kirjutada igas situatsioonis kasutamiskõlbliku valemi kahepoolse testi p-väärtuse leidmiseks:

$$\text{p-väärtus} = 2 \cdot \Phi\left(-\frac{|c - 0,5n(N+1)| - 0,5}{\sqrt{mn(N+1)/12}}\right).$$

3.4 Populatsiooni mudelist

Wilcoxon astaksumma testi rakendamiseks pole meil tarvis endale ette kujutada kahte populatsiooni vms. Piisab sellest, et me jagame uuritavad juhuslikult kahte gruppi. Wilcoxon testi puhul saame pärida, kas antud uuritavate indiviidide puhul üks töötlus on parem kui teine töötlus.

Enamasti aga soovime me teha järeldusi mitte meie valimisse kuuluvate indiviidide kohta, vaid mingi laiema populatsiooni kohta. Näiteks võime me võtta valimi patsientidest, raviga osasid ühel ja teisel viisil ning lõpuks soovime öelda, milline ravi on parem antud haiguse patsientidele (kõigile patsientidele, mitte ainult meie valimisse sattunud inimeste jaoks). Või võtame kaks valimit kahest populatsioonist, $X \sim F(x)$ ja $Y \sim G(y)$ ning soovime teada, kas mõlemas populatsioonis on uuritava tunnuse jaotus samasugune, kas $F = G$?

Võtame valimi suurusega $N = n + m$. Kui nullhüpotees kehtib, $F = G$, siis on vaatluste $X_1, \dots, X_n, Y_1, \dots, Y_m$ mistahes järjestus võrdvõimalik. Seega on tõenäosus näha esimeses valimis (X -id) astakuid $s_1, \dots, s_n$ on võrdne võimaluste arvuga valida N vaatluse seast välja n vaatlust,

$$P(s_1, \dots, s_n) = 1/C_N^m.$$

Seega jääb nullhüpoteesi kehtides Wilcoxon astaksummatesti statistiku jaotus samasuguseks, ükskõik kas meil on tegemist juhuslike valimitega populatsioonidest või me kõigest jagame juhuslikult gruppidesse kogu meid huvitavat populatsiooni.

Mõnevõrra segadust võib tekkida siis, kui uuritava tunnuse jaotus pole (päriselt) pidev. Sellisel juhul on igas valimis erinev arv võrdseid vaatluseid ja Wilcoxon astaksummatesti teststatistiku jaotus jääb sõltuma uuritavast jaotusest (sellest, millise tõenäosusega võime näha kokkulangevaid vaatluseid). Kuna aga uuritava tunnuse tegelik jaotus pole ju teada (milleks muidu me kasutame mittepameetrist testi), siis pole esmapilgul võimalik p-väärtust arvutada ja reaalselt statistilist testi teha. Lahenduseks on nn tingliku testi kasutamine — eeldame, et võrdsete astakute konfiguratsioon on antud, ja muretseme siis, et sellise konkreetse võrdsete astakute konfiguratsiooni puhul I-liiki viga ei tehtaks lubatust suurema tõenäosusega. Olulisustõenäosuse arvutamine toimub sellisel juhul täpselt samal viisil, kui tegime seda randomiseeritud katse korral. Kuna mistahes võrdsete astakute konfiguratsiooni puhul me ei tee lubatust sagedamini esimest liiki viga, siis võime olla kindlad, et selline testprotseduur garanteerib, et I-liiki vea tegemise tõenäosus poleks suurem kui uurija poolt valitud olulisuse nivoo α .

3.5 Testi võimsus

Kas me suudame oma kasinate võimete ja võimaluste juures tõestada alternatiivse hüpoteesi (kui alternatiivne hüpotees ikka on õige)? Kui suurt valimit läheb tarvis, et suudaksime näidata erinevust populatsioonide vahel?

Üldjuhul käsitleb alternatiivne hüpotees ($F \neq G$) sedavõrd paljusid erinevaid situatsioone, et midagi üldist kõigi nende olukordade kohta öelda on raske. Kui uurijal on silma ees mingi konkreetne ettekujutus sellest, kuidas kahe populatsiooni jaotused võiksid teineteisest erineda, siis on sageli kõige lihtsam tee arvutisimulatsiooni abil hinnata testi võimsust (genereerida k sellist valimit, kus n vaatlust on võetud jaotusest F ja m vaatlust jaotusest G ja vaadata siis, kui paljudes genereeritud valimites suutis astaksummatest vastu võtta alternatiivse hüpoteesi).

Mida suuremaks muutuvad valimi suurused, seda raskemaks läheb ka nende genereerimine. Sestap vaatame alljärgnevalt mõnda võimalust leida astaksummatesti asümptootilist võimsust. Järgnevates arvutustes on veidi mugavam kasutada teststatistiku W_s asemel nn Mann-Whitney teststatistikut W_{xy} .

Alustame sellest, et leiame teststatistiku keskväärtuse ja dispersiooni alternatiivse hüpoteesi kehtides.

3.5.1 Keskvärtus ja dispersioon H_1 kehtides

Olgu antud n sõltumatut sama jaotusega juhuslikku suurust jaotusest F — $X_1, X_2, \dots, X_n \sim F$ ja m s.s.j.j.s. jaotusest G — $Y_1, Y_2, \dots, Y_m \sim G$.

Defineerime uued juhuslikud suurused U_{ij} :

$$U_{ij} = \begin{cases} 1 & , \text{ kui } X_i > Y_j \\ 0 & \text{ muidu} \end{cases}.$$

Juhusliku suuruse U_{ij} keskväärtuseks on

$$EU_{ij} = P(X_i > Y_j) =: p_1.$$

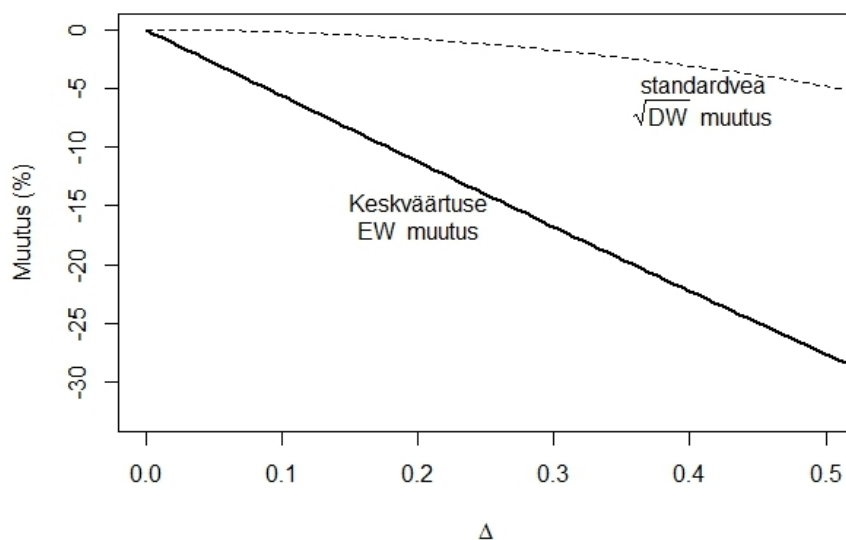
Mann-Whitney teststatistik aga on esitatav juhuslike suuruste U_{ij} kaudu:

$$W_{xy} = \sum_i \sum_j U_{ij}.$$

Seega $EW_{xy} = \sum_i \sum_j EU_{ij} = mnp_1$. Siit saame ühe täiendava võimaluse interpreteerida teststatistiku väärtust — nimelt on suurus $W_{xy}/(mn)$ nihketa hinnanguks tõenäosusele $p_1 = P(X > Y)$.

Mis juhtub teststatistiku dispersiooniga juhul, kui kehtib alternatiivne hüpotees? Kui erinevus jaotuste F ja G vahel on väike, siis muutub ka Mann-Whitney teststatistiku dispersioon (või standardviga) vaid veidi. Enamasti märkimisväärselt vähem kui teststatistiku keskvärtus. Kui näiteks võrreldavad populatsioonid oleksid normaaljaotusega, keskvärtuste erinevusega Δ , siis muutust Mann-Whitney teststatistiku keskvärtuses ja dispersioonis kirjeldab joonis 3.1. Näeme, et väikeste Δ väärtuste puhul on muutus test-

Joonis 3.1: Teststatistiku W_{xy} keskvärtuse ja standardvea protsentuaalne muutus, kui keskvärtuste erinevus on Δ ja uuritava tunnuse jaotuseks on normaaljaotus.



statistiku dispersioonis väga väike, ignoreeritav. Paljud autorid ütlevadki, et kui F ja G on lähedased, siis $DW_{xy} \approx D_0 W_{xy}$. Ka antud konspektis on hiljem kasutatud seda ligikaudset võrdust. Siinkohal paneme aga kirja ka valemi teststatistiku dispersiooni täpseks arvutamiseks.

$$DW_{xy} = \sum_i \sum_j DU_{ij} + \sum_i \sum_j \sum_k \sum_l \text{Cov}(U_{ij}, U_{kl}),$$

kus kovariatsioonide summeerimine toimub üle nende i, j, k ja l väärtuste, mille puhul kas $i \neq k$ ja/või $j \neq l$ (variandi $i = k$ ja $j = l$ puhul on tegemist U_{ij} dispersiooniga).

Antud summas on kolme erinevat tüüpi kovariatsioone. Kui $i \neq k$ ja $j \neq l$,

siis leitakse U_{ij} ja U_{kl} täiesti erinevaid X ja Y tunnuse väärtuseid kasutades ja on sõltumatud (kuna on sõltumatute juhuslike suuruste funktsioonid). Seega on vastavad kovariatsioonid kõik nullid.

Juhtumeid, kus $i = k$ aga $j \neq l$ on $nm(m-1)$ tükki (X -e on n tükki, j valikuks m võimalust Y -te seast ja l valikuks jääb järgi $m-1$ võimalust). Liidetavaid, kus $i \neq k$ aga $j = l$ on aga $nm(n-1)$ tükki. Seega

$$DW_{xy} = mnDU_{ij} + nm(m-1)\text{Cov}(U_{ij}, U_{il}) + nm(n-1)\text{Cov}(U_{ij}, U_{kj}). \quad (3.3)$$

Leiame valemis kirjas olevad dispersioonid ja kovariatsioonid. Kuna U_{ij} on binaarne, siis

$$DU_{ij} = p_1(1 - p_1).$$

Kuna $E(U_{ij}U_{il}) = P(U_{ij} = U_{il} = 1) = P(X_i > Y_j, X_i > Y_l) =: p_2$ siis

$$\begin{aligned} \text{Cov}(U_{ij}, U_{il}) &= E(U_{ij}U_{il}) - EU_{ij} \cdot EU_{il} \\ &= p_2 - p_1^2. \end{aligned}$$

Tähistades $p_3 := P(X_i > Y_j, X_k > Y_j)$ saame analoogselt, et

$$\text{Cov}(U_{ij}, U_{kj}) = p_3 - p_1^2.$$

Asendades saadud tulemused valemisse 3.3 saame

$$DW_{xy} = mnp_1(1 - p_1) + mn(m-1)(p_2 - p_1^2) + nm(n-1)(p_3 - p_1^2).$$

Juhul kui H_0 kehtib, siis $p_1 := P(X > Y) = 1/2$ ja $p_2 = p_3 = 1/3$ (juhuslikult valitud vaatluse tõenäosus olla väiksem kolmest sama jaotusega vaatlusest). Sellisel juhul aga $E_0 W_{xy} = mnp_1 = mn/2$ ja

$$\begin{aligned} D_0 W_{xy} &= mn/4 + nm(m-1)(1/3 - 1/4) + nm(n-1)(1/3 - 1/4) \\ &= mn(3 + m - 1 + n - 1)/12 \\ &= mn(n + m + 1)/12 \\ &= mn(N + 1)/12. \end{aligned}$$

Saadud tulemused langevad kokku varem nähtud tulemustega.

3.5.2 Asümptootiline võimsus

Kontrollime nullhüpoteesi $H_0 : G = F$ kehtivust. Oletame, et nullhüpotees lükatakse ümber, kui teststatistik osutub suuremaks kriitilisest väärtusest

c_{krit} , $W_{xy} \geq c_{krit}$ (vaatame ühepoolset hüpoteesi). Siis alternatiivse hüpoteesi ($F < G$) kehtides on testi võimsus $\Pi(F, G)$ leitav valemist:

$$\begin{aligned}\Pi(F, G) &= P(W_{xy} \geq c_{krit} | H_1) \\ &= P\left(\frac{W_{xy} - mnp_1}{\sqrt{D_1 W_{xy}}} \geq \frac{c_{krit} - mnp_1}{\sqrt{D_1 W_{xy}}} | H_1\right) \\ &\approx 1 - \Phi\left(\frac{c_{krit} - mnp_1}{\sqrt{D_1 W_{xy}}}\right),\end{aligned}\quad (3.4)$$

sest

$$\frac{W_{xy} - mnp_1}{\sqrt{D_1 W_{xy}}} \xrightarrow{\mathcal{D}} N(0, 1).$$

Teades p_1 , p_2 ja p_3 väärtuseid, saame leida testi võimsuse. Nimetatud tõenäosuste leidmine on vahel võimalik, aga sageli on just p_2 ja p_3 jaoks millegi mõistliku väljapakkumine praktikas keerukas. Sestap proovime leida lisaeeldusi ja täiendavaid lihtsustusi tehes praktikas lihtsamini kasutatava valemi.

Esmalt leiame, milline on kriitiline väärtus c_{krit} :

$$\begin{aligned}\alpha &= P(W_{xy} \geq c_{krit} | H_0) \\ \alpha &= P\left(\frac{W_{xy} - nm/2}{\sqrt{D_0 W_{xy}}} \geq \frac{c_{krit} - nm/2}{\sqrt{D_0 W_{xy}}} | H_0\right) \\ \alpha &\approx 1 - \Phi\left(\frac{c_{krit} - nm/2}{\sqrt{D_0 W_{xy}}}\right) \\ z_{1-\alpha} &\approx \frac{c_{krit} - nm/2}{\sqrt{D_0 W_{xy}}}\end{aligned}$$

Kus $z_{1-\alpha}$ on standardse normaaljaotuse $(1 - \alpha)$ -kvantiil. Paari lihtsa teisendusega (ja arvestades, et $D_0 W_{xy} = mn(N + 1)/12$) jõuame soovitud tulemuseni:

$$c_{krit} = z_{1-\alpha} \sqrt{mn(N + 1)/12} + nm/2.$$

Asendades saadud tulemuse võimsuse valemisse 3.4 saame:

$$\begin{aligned}\Pi(F, G) &= 1 - \Phi\left(\frac{z_{1-\alpha} \sqrt{mn(N + 1)/12} + nm/2 - mnp_1}{\sqrt{D_1 W_{xy}}}\right) \\ &= 1 - \Phi\left(\frac{nm(0.5 - p_1) - z_{\alpha} \sqrt{mn(N + 1)/12}}{\sqrt{D_1 W_{xy}}}\right).\end{aligned}\quad (3.5)$$

Saadud valemit on võimalik mõnikord tõepoolest kasutada. Näiteks eeldades, et uuritava tunnuse jaotus mõlemas populatsioonis on normaaljaotus, sama hajuvusega (standardhälbega σ) ning eeldades, et keskväärtuste erinevus populatsioonide vahel on Δ , saame tõenäosused p_1, p_2, p_3 ja ühepoolse testi võimsuse leida näiteks järgmise R programmi abil:

```
# Valimite suurused
n=30; m=10

# Uuritava tunnuse hajuvus (mõlemas populatsioonis)
sigma=1

# Keskväärtuste erinevus
delta=0.5

# Kasutatav olulisuse nivoo
alpha=0.05

# Leiame tõenäosused p1, p2, p3 (juhul, kui on normaaljaotus)
library(mvtnorm)
p1=pnorm(0, -delta, sd=sqrt(2*sigma**2))
p2=pmvnorm( lower=c(-Inf,-Inf), upper=c(0,0),
             mean=c(-delta, -delta),
             sigma=cbind(c(2*sigma**2,sigma**2), c(sigma**2,2*sigma**2)))

# Teststatistiku dispersioon ja keskväärtus H1 kehtides:
DW=m*n*p1*(1-p1)+m*n*(n-1)*(p2-p1**2)+ n*m*(m-1)*(p3-p1**2)
EW=m*n*p1

# Leiame ühepoolse testi võimsuse
1-pnorm((n*m*(0.5-p1)-qnorm(alpha)*sqrt(m*n*(m+n+1)/12))/sqrt(DW))
```

Toodud valemit otse kasutada on enamasti siiski raske. Paremal juhul võib eeldada, et $DW_{xy} \approx D_{H_0}W_{xy}$, millisel juhul läheb võimsusarvutuse jaoks tarvis vaid suurust $p_1 := P(X > Y)$. Ka selle ühe tõenäosuse leidmine ei pruugi praktikas olla kuigi lihtne ülesanne. Milline võiks näiteks olla tõenäosus, et hommikuti putru söönud laps on pikem putru mitesöönud lapsest? Palju lihtsam on aga saada vanaemalt kätte eksperthinnang: tänu pudrule kasvab laps 5cm pikemaks. Vaatamegi alljärgnevalt, kuidas kasutada valemit 3.5 juhul, kui teame testitava töötluse kohta, palju ta uuritava tunnuse väärtust muudab.

Erinegu kaks populatsiooni teineteisest vaid nihke poolest, st $Y = X - \Delta$ ehk $F_X(x) = G_Y(x - \Delta)$. Sellisel juhul $p_1 = P(X - Y > 0) = P(X - (Y + \Delta) < -\Delta)$. Kuna nii X kui ka $Y + \Delta$ on sama jaotusega, siis $p_1 = 1 - F^*(-\Delta) = F^*(\Delta)$, kus F^* on kahe F jaotusega juhusliku suuruse vahe jaotus.

Funktsiooni $f(x)$ Tayloriga rida kohal a on kujul

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x - a)^n.$$

Tähistades F^* tihedusfunktsiooni f^* -ga saame kirja panna funktsiooni F^* Tayloriga rittaarenduse kohal 0:

$$\begin{aligned} F^*(\Delta) &= F^*(0) + f^*(0)\Delta + \frac{(f^*)'(0)}{2}\Delta^2 + \dots \\ &\approx F^*(0) + f^*(0)\Delta \end{aligned}$$

Kuna $F^*(0) = 0.5$ — sest kahe sama jaotusega juhusliku suuruse vahe mediaan (ja keskväärus) on 0 — siis saame, et

$$p_1 - 0.5 \approx \Delta f^*(0).$$

Asendame saadud tulemuse valemisse 3.5 ja saame:

$$\begin{aligned} \Pi(F, G) &= 1 - \Phi \left(\frac{-mn\Delta f^*(0) - z_\alpha \sqrt{\frac{1}{12}mn(N+1)}}{\sqrt{D_1 W_{xy}}} \right) \\ &= \Phi \left(\frac{mn\Delta f^*(0) + z_\alpha \sqrt{\frac{1}{12}mn(N+1)}}{\sqrt{D_1 W_{xy}}} \right), \end{aligned}$$

sest $1 - \Phi(x) = \Phi(-x)$. Kui Δ on väike, siis $D_1 W_{xy} \approx \frac{1}{12}mn(N+1)$ ja

$$\Pi(F, G) = \Phi \left(\sqrt{\frac{12mn}{N+1}} \Delta f^*(0) + z_\alpha \right).$$

Näide 3.1 *Küsitlusfirmas töötab 20 ankeedisestajat kahes tööruumis. Meid huvitab, kas taustamuusika mängimine tööruumis võiks tõsta nende töö efektiivsust. Teame varasemast, et ühe töötaja poolt päeva jooksul sisestatud lehekülgede arvu dispersioon on 32. Kui suure tõenäosusega suudaksime tõestada olulisuse nivool 0,05 taustamuusikast tingitud muutuse (eeldame hetkel, et*

meid huvitab vaid ühepoolne hüpotees)? Oletame, et kui muutus eksisteerib, siis tõstab taustamuusika sisestatud lehekülgede arvu 5 võrra. Soovime kasutada Wilcoxon testi (sest me ei tea kindlalt päeva jooksul sisestatud lehekülgede arvu jaotust), aga võimsusarvutust tehes eeldagem, et vaadeldud tunnus on normaaljaotusega.

Lahendus.

Kui $X, X' \sim N(\mu, \sigma^2 = 32)$ siis $X - X' \sim N(0, 2 \cdot \sigma^2)$ ja seega $F^* = N(0, 2 \cdot \sigma^2)$ ning

$$\begin{aligned} f^*(x) &= \frac{1}{\sqrt{2\pi(2 \cdot \sigma^2)}} \exp\left(-\frac{(x - \mu)^2}{2\sqrt{(2 \cdot \sigma^2)}}\right) \\ f^*(0) &= \frac{1}{2\sigma\sqrt{\pi}}. \end{aligned}$$

Seega oleks Wilcoxon astaksummatesti võimsus $\Delta = 5$ korral

$$\begin{aligned} \Pi(F, G) &= \Phi\left(\sqrt{\frac{12mn}{N+1}} \frac{\Delta}{2\sigma\sqrt{\pi}} + z_\alpha\right) \\ &= \Phi\left(\sqrt{\frac{12 \cdot 10 \cdot 10}{20+1}} \frac{5}{2\sqrt{32}\sqrt{\pi}} - 1,64\right) \\ &= \Phi(0,24) \\ &= 0,5948. \end{aligned}$$

Testi võimsus 0,59 on madalapoolne, aga riskialtimate eksperimentaatori jaoks ehk siiski piisav katsega alustamiseks. Kui leiaksime asümptootilise võimsuse abil testi täpse võimsuse, siis see oleks 0,55 (valimimaht 10+10 pole veel täiesti piisav asümptootika probleemivabaks tööks, aga väga viltu nüüd ka ei läinud...).

Vahel soovitakse testi võimsuse asemel leida hoopis valimi suurust, mida oleks vaja teatud võimsuse saavutamiseks. Valimi mahu leidmiseks, mis tagaks testi võimsuse β , peaksime leidma m ja n väärtused järgmisest võrrandist:

$$\beta = \Phi\left(\sqrt{\frac{12mn}{N+1}} \Delta f^*(0) + z_\alpha\right).$$

Saamaks ühest lahendist tehakse mingi eeldus m ja n soovitava suhte kohta (võib fikseerida ka ühe neist ja küsida, kui suure valimi peaksime võtma teisest populatsioonist...). Näiteks võime otsustada, et $m = n$. Sellisel juhul (ja arvestades, et $1/(N+1) \approx 1/N$ suure valimi korral):

$$\begin{aligned}
\beta &= \Phi \left(\sqrt{\frac{12mn}{N+1}} \Delta f^*(0) + z_\alpha \right) \\
z_\beta &= \sqrt{\frac{12mn}{N+1}} \Delta f^*(0) + z_\alpha \\
z_\beta - z_\alpha &= \sqrt{\frac{12n^2}{2n}} \Delta f^*(0) \\
\frac{z_\beta - z_\alpha}{\Delta f^*(0)} &= \sqrt{6n} \\
\frac{(z_\beta - z_\alpha)^2}{\Delta^2 (f^*(0))^2} &= 6n \\
n &= \frac{(z_\beta - z_\alpha)^2}{6\Delta^2 (f^*(0))^2}
\end{aligned}$$

Näide 3.2 Pöördume tagasi eelmise, ankeedisisestajate näite juurde. Vaatame, kui suurt valimit läheks meil vaja, et saavutada 90% võimsust:

$$\begin{aligned}
n &= \frac{(z_\beta - z_\alpha)^2}{6\Delta^2 \frac{1}{4\sigma^2\pi}} \\
&= (z_\beta - z_\alpha)^2 \frac{\sigma^2}{\Delta^2} 2\pi/3 \\
&= (1,28\dots - (-1,64\dots))^2 \cdot \frac{32}{5^2} \cdot 2 \cdot 3,14\dots/3 \\
&= 22,95.
\end{aligned}$$

Mõlemasse katserühma vajame seega 23 ankeedisisestajat.

3.5.3 Wilcoxon ja t-testi võrdlus

Paljudel tekib kindlasti küsimus — kui palju me võidaksime testi võimsuses, kui kasutaksime Wilcoxon testi asemel t-testi? Erinevate testide võrdlemisel kasutatakse sageli nn *Pitman efficiency*-t ehk Pitmani efektiivsust (vahel kutsutakse ka relatiivseks efektiivsuseks ehk *relative efficiency*). Nimetatud näitaja leidmiseks tuleb välja selgitada, kui suurt valimit läheb vaja ühe testi korral ja kui suurt valimit teise testi korral et saavutada etteantud võimsus. Kahe valimi suuruse suhe ongi Pitmani efektiivsus. Näiteks Wilcoxon ja t-testi puhul oleks Pitmani efektiivsus $e_{t,W}$ leitav suhtena

$$e_{t,W} = \frac{n_t}{n_{Wilcoxon}},$$

kus n_t oleks valimi suurus, mida t-test vajab võimsuse β saavutamiseks ja $n_{Wilcoxon}$ on Wilcoxon testi poolt vajatav valimi suurus sama testi võimsuse korral. Arvutame t-testi suhtelise efektiivsuse Wilcoxon testi suhtes suurte valimite jaoks (kus võime kasutada teststatistikute asümptootilisi jaotusi). Asümptootilise Relatiivse Efektiivsuse tähistamiseks kasutatakse sageli lühendit ARE.

Sarnast tuletuskäiku (kui Wilcoxon testi puhul) kasutades, võime jõuda t-testi puhul vajamineva valimi mahu suuruse arvutamisel tulemuseni

$$n_t = 2(z_\beta - z_\alpha)^2 \frac{\sigma^2}{\Delta^2}.$$

Juhul, kui vaatlused oleksid normaaljaotusega, siis oleks t-testi ARE Wilcoxon astaksumma testi suhtes seega:

$$\begin{aligned} e_{t,W} &= \frac{2(z_\beta - z_\alpha)^2 \cdot \sigma^2 / \Delta^2}{(z_\beta - z_\alpha)^2 \cdot \sigma^2 / \Delta^2 \cdot \frac{2\pi}{3}} \\ &= \frac{3}{\pi} \\ &\approx 0,955 \end{aligned}$$

Seega justnagu "kaotaksime" Wilcoxon testi kasutades umbes 5% oma valimi mahust. Samas võime teha sarnaseid arvutusi ka teiste jaotuste korral. Kui valim on suur, võib ju t-testi kasutada ka siis, kui vaatlused pole normaaljaotusega. Alljärgnevalt mõned ARE-numbrid erinevate jaotuste puhul, nihkemudel:

normaaljaotus	logistiline	eksponentjaotus	double exponent
0,955	1,097	3	1,5

Enamike mittenormaalsete jaotuste puhul on Wilcoxon test tundlikum ja võib vahel vajada isegi 3-4 korda väiksemat valimit kui t-test. Eksisteerib ka alampiiir — t-testi ja Wilcoxon testi relatiivne efektiivsus ei saa (nihkemudeli korral suvalise pideva jaotuse F puhul) olla väiksem kui 0,864 ($e_{t,W} \geq 0,864$). Vaata ka joonist 3.2. Vastavat ülempiiri ei eksisteeri — Wilcoxon test võib olla kuidas palju tundlikum t-testist. Seega võimalik võit Wilcoxon astaksummatesti kasutades võib olla väga suur, aga kaotus (kui vaatlused näiteks tõepoolest on normaaljaotusega) üsnagi pisike.

Märkus: samasugune alampiiir kehtib ka Wilcoxon astakmäärgitesti korral (võrreldes sõltuvate valimite jaoks mõeldud t-testiga).

3.6 Wilcoxon astaksummatesti tulemuse tõlgendamisest

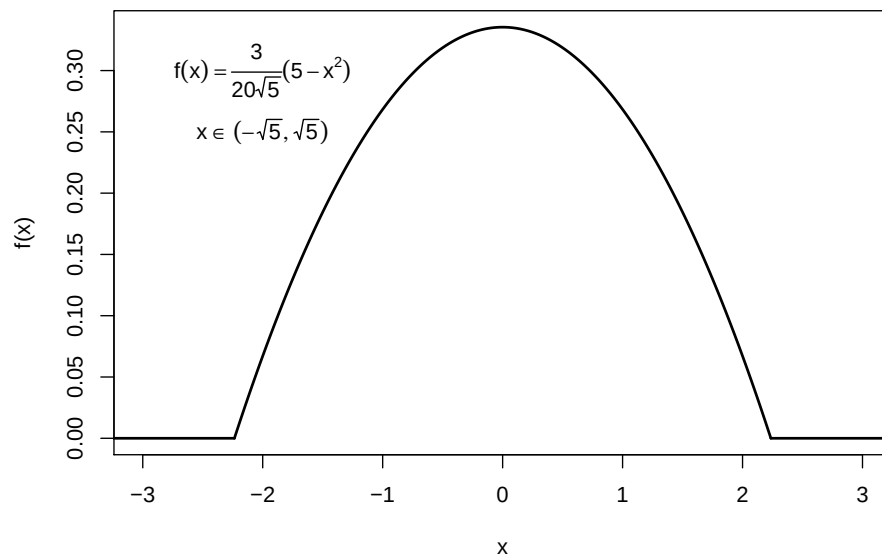
Kui saame Wilcoxon astaksummatesti abil lugeda tõestatuks alternatiivse hüpoteesi, siis mida see ikkagi tähendab? Kirjanduses eksisteerib väga palju eksitavaid väiteid selle kohta, mida Wilcoxon astaksummatest siis ikkagi kontrollib. Nii nagu leidub eksitavaid väiteid ka selle kohta, millised on Wilcoxon astaksummatesti eeldused. Näiteks kohtab väiteid, mis väidavad, et Wilcoxon astaksummatest kontrollib keskväärtuste või mediaanide erinevust.

See väide ei pea paika.

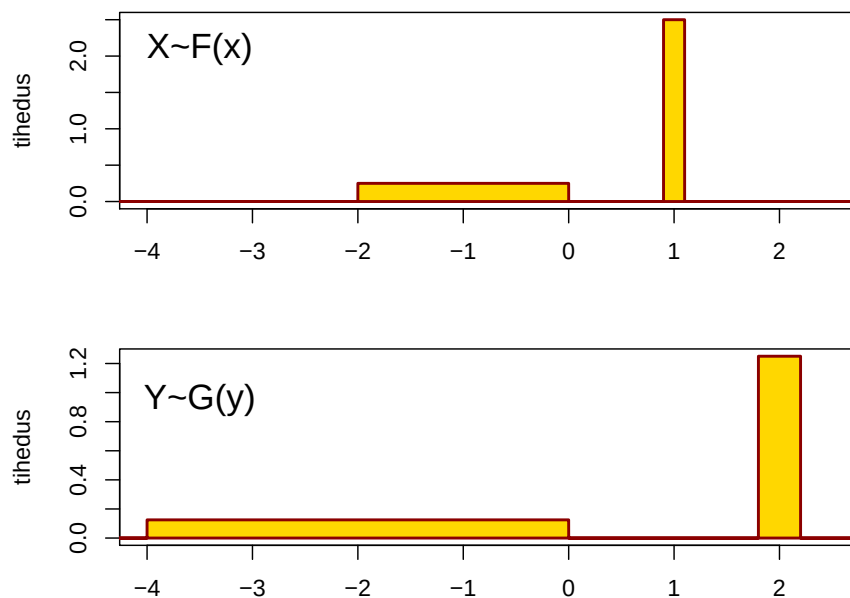
Saab tuua näite olukorrast, kus mõlemas kontrollitavas populatsioonis on nii uuritava tunnuse keskväärtused kui ka mediaanid võrdsed, kuid Wilcoxon astaksummatest võtab kindlalt vastu alternatiivse hüpoteesi (vähegi mõistlikuma valimisuuruse juures). Vaata näiteks joonist 3.3.

Asjakohasem oleks ikkagi öelda: kui võtame ühest populatsioonist ühe juhuslikult valitud esindaja X ja teisest populatsioonist ühe juhuslikult valitud esindaja Y , siis Wilcoxon astaksummatest kummutab nullhüpoteesi siis, kui $p_1 := P(X > Y) \neq 0,5$. Ülaloodud näite korral on $p_1 \approx 0,56$, seega

Joonis 3.2: Jaotus, mille korral $e_{t,W} = 0,864$ (maksimaalselt t-testi soosiv jaotus).



Joonis 3.3: Kaks jaotust, mille võrdlemisel Wilcoxon astaksummatest võtab vastu alternatiivse hüpoteesi, kuigi mõlema jaotuse mediaan ja keskvärtus on samad.



on ka alternatiivse hüpoteesi vastuvõtmine Wilcoxon testi poolt õigustatud. Vahel mainitaksegi, et hüpotees mida Wilcoxon astaksummatest kontrollib, on väide $H_0 : P(X > Y) = 0,5$.

Samuti võib segadustekitavalt mõjuda mõnede allikate väide, et Wilcoxon testi tohib kasutada vaid siis, kui $F(x) = G(x + \Delta)$ ehk nn nihkealternatiivi korral. Kui tegemist on jaotuse nihkega, siis saab Wilcoxon astaksummatesti tõepoolest interpreteerida kui testi mediaanide (või keskvärtuste, kui keskvärtus eksisteerib) võrdlemiseks. Paraku on väga raske kindlaks teha, kas vastav tingimus on täidetud. Samuti on ta suuresti mittevajalik — kui antud tingimus pole täidetud, ei juhtu Wilcoxon testiga midagi muud, kui vaid see, et me ei tohi enam antud testi interpreteerida kui testi mediaanide (või keskvärtuste) võrdlemiseks.

Kui Wilcoxon astaksummatest võtab vastu alternatiivse hüpoteesi, siis efekti suurust saab iseloomustada kasutades hinnangut suurusele $P(X > Y)$, sobivaks hinnanguks oli $W_{xy}/(m \cdot n)$.

3.7 Võrdsed vaatlused

Kui leidub vaatluseid, mida pole võimalik järjestada, kasutatakse taas keskmiseid astakuid ehk *midrank*'e. Loomulikult muutub seetõttu ka Wilcoxon astaksummatesti teststatistiku jaotus (näiteks võib tekkida võimalus näha mittetäisarvulist teststatistiku väärtust). Kui leiame teststatistiku jaotust kõigi võimaluste läbivaatamise teel, pole sellest suuremat häda (tuleb vaid mees pidada, et näiteks keskmiste astakute 3,5 ja 3,5 puhul on tegemist kahe erineva numbriga, seega astakute kombinatsioonid $\{1; 3,5_1; 7\}$ ja $\{1; 3,5_2; 7\}$ on kaks erinevat kombinatsiooni).

Kui soovime kasutada teststatistiku asümptootilist jaotust, siis tuleb mees pidada, et teststatistiku dispersioon muutub tänu keskmiste astakute kasutamisele:

$$D_0 W_{xy}^* = D_0 W_s^* = \frac{mn(N+1)}{12} - \frac{mn \sum_i^e (d_i^3 - d_i)}{12N(N-1)},$$

kus e on erinevate väärtuste arv ühendatud valimis ja d_i on i . unikaalse väärtuse korduste arv (mitu korda seda väärtust esines kahe võrreldava valimi peale kokku).

Märkus: statistiku W_s^* (või W_{xy}^*) jaotus ei pruugi olla normaaljaotusele lähedane isegi suure valimi korral, kui $\max_i d_i/N \approx 1$ ehk kui mõni konkreetne uuritava tunnuse väärtus domineerib mõlemas valimis.

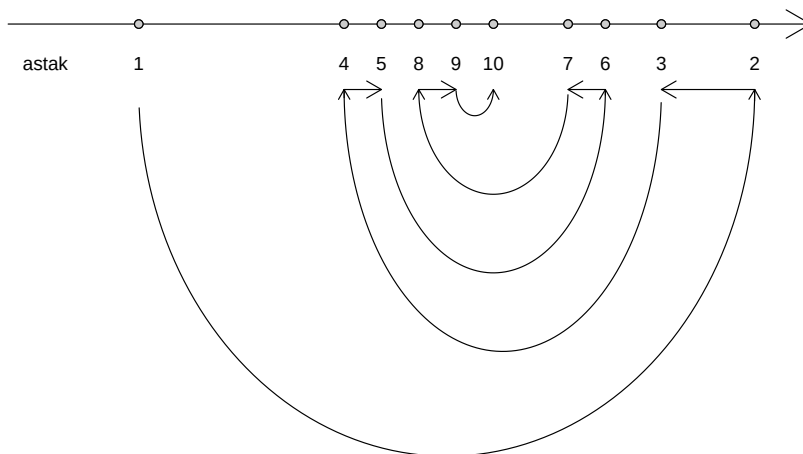
Märkus 2: pidevuse paranduse kasutamine on keeruline, kui valimites esineb samasuuri uuritava tunnuse väärtuseid — teststatistiku võimalike väärtuste vahelised kaugused hakkavad sellises olukorras varieeruma. Seega $+0,5$ ja $-0,5$ asemel võib tarvis minna ka $+0,25$ või $-0,25$ või Ühesõnaga, võrdse suurusega vaatluseid sisaldava andmestiku korral pidevuse parandust enamasti (praktilistel põhjustel) ei kasutata.

3.8 Siegel-Tukey ja Ansari-Bradley testid

Wilcoxon testi puhul on alternatiivne hüpotees $F \neq G$ väga üldine. Osade konkreetsete alternatiivide korral on Wilcoxon testi võimsus madal. Näiteks kuigi $F \neq G$, aga kui $P(X > Y) \approx 0,5$, siis Wilcoxon testi võime alternatiivse hüpoteesi kehtivust tõestada on väga madal. Selline situatsioon võib näiteks tekkida siis, kui kaks võrreldavat populatsiooni erinevad eelkõige selle poolest, et ühes populatsioonis on uuritava tunnuse hajuvus suurem kui teises (suur geneetiline varieeruvus metsikus populatsioonis vs loomaialloomade või aretatud taimesortide ühetaolisus; rikkis kontrollaparatuur ei suuda

tagada toodangu ühetaolisust jne). Parandamaks testi võimet eristada just nimelt erinevust varieeruvuses, võime muuta astakute paneku järjekorda. Väikseim väärtus saab astaku 1, suurim väärtustest saab astaku 2, suuruselt teine vaatlus saab astaku 3 jne, vaata ka joonist 3.4. Teststatistikuks on sarnaselt Wilcoxon astaksummatestile ühe valimi vaatluste astakute summa. Saadud statistiku jaotus nullhüpoteesi kehtides on sama mis Wilcoxon testil.

Joonis 3.4: Astakute jaotamine Siegel-Tukey testi korral



Siegel-Tukey testile heidetakse ette, et testi statistiku väärtus (ja testi tulemus) võib sõltuda sellest, kummast otsast (suurimast või väikseimast väärtusest) astakute jagamist alustatakse. Variatsioonrea mõlemaid otsi kohtleb poliitilisemalt korrektsemalt (võrdsemalt) Ansari-Bradley test. Nimetatud testi puhul antakse nii suurimale kui ka väiksemale väärtusele astak 1, variatsioonrea 2. ja eelviimasele vaatlusele astak 2 jne. Taolisel viisil saadud astakute summa ühe grupi jaoks annabki Ansari-Bradley teststatistiku väärtuse. Kahjuks pole aga Ansari-Bradley teststatistiku jaotus nullhüpoteesi kehtides aga sama kui Wilcoxon testil.

Märkus: Siegel-Tukey ja Ansari-Bradley testide võime tuvastada hajuvuse erinevust kahaneb, kui lisaks erinevatele hajuvustele on ühes populatsioonis uuritava tunnuse väärtused suuremad kui teises. Üks praktikas kasutatav lahendus nimetatud probleemile oleks järgmine: hinnatakse Wilcoxon teststatistikut kasutades võimalikku nihet Δ , mille võrra tuleks populatsioone

nihutada, et $P(X > Y) = 0.5$. Tehaksegi soovitud nihe, $Y_{uus} = Y - \Delta$. pärast tehakse siis kas Siegel-Tukey või Ansari-Bradley test nihutatud väärtuseid kasutades. Nimetatud testprotseduur pole päris korrektne, aga suure valimi puhul on võimalikud probleemid on tühised ja ignoreeritavad.

Näide 3.3 Kahtlustatakse, et teatud ainega X kokkupuutel kaob inimorganismi võime kontrollida β -glükoosi taset veres. Aga β -glükoosi taseme järsud muutused võivad kahjustada maksa ja põhjustada ajapikku tõsiseid probleeme. Otsustatakse testida, kas ainega X kokkupuutunud inimestel on β -glükoosi kontsentratsioon varieeruvam kui kontrollgrupi inimestel.

<i>Eksponeeritud</i>	67, 80, 120
<i>Kontrollgrupp</i>	78, 82, 84, 86, 90, 92, 95

Siegel-Tukey teststatistik on $W_{ST} = 1 + 2 + 5 = 8$. Veel väiksema teststatistiku väärtuse saamise võimalusi on 4: $\{1, 2, 3\}$; $\{1, 2, 4\}$; $\{1, 2, 5\}$; $\{1, 3, 4\}$. Kokku on võimalusi näpsata 3-e astakut 10 seast $C_{10}^3 = 120$. Ühepoolse testi olulisustõenäosus on seega $4/120 = 0,0333\dots$ ja kahepoolse testi olulisustõenäosus on $0,0666\dots$

Ansari-Bradley testi korral oleks teststatistiku väärtuseks $W_{AB} = 1 + 1 + 3 = 5$, astakute komplekte, mis annaksid samasuure või väiksema statistiku väärtuse, oleks 6 (kuna iga astakut esineb kaks korda, siis eristame neid alaindeksitega v ja p): $\{1_v, 1_p, 2_v\}$; $\{1_v, 1_p, 2_p\}$; $\{1_v, 1_p, 3_v\}$; $\{1_v, 1_p, 3_p\}$; $\{1_v, 2_v, 2_p\}$; $\{1_p, 2_v, 2_p\}$. Ühepoolse testi olulisustõenäosus tuleks seega $6/120 = 0,05$ ja kahepoolse testi korral $0,1$.

Ülesanded

1. Kõige raskemate, viimases staadiumis vähihaigete peal katsetati eksperimentaalset ravimeetodit. Osad patsientidest randomiseeriti eksperimentaalset ravi saavasse patsientide gruppi, osad tavapärast ravi saavasse kontrollgruppi. Patsientide elulemused (kuudes) mõlemas grupis olid järgmised:

Eksperimentaalne grupp:

28.6 11.4 7.9 10.7 6.1 5.6 10.8 44.6 16.7 15.0 19.4 5.4 8.2 18.2 7.0 15.2
8.0 12.3 12.5 7.4 15.8 15.3 17.9 17.5 10.5

Kontrollgrupp:

7.6 11.8 1.5 1.4 4.4 28.9 12.3 5.4 9.6 1.5 13.9 7.6 12.4 44.2 10.5 10.4
18.8 6.5 3.4 5.9

Leia Wilcoxon'i astaksummatesti abil, kas elukestvuse jaotus on erinev nendes kahes patsientide grupis. Kui näed erinevust, siis milline see erinevus on (kumb ravimeetod töötas paremini?). Võrdle ravimeetodeid ka t-testi abil, kommenteeri saadud tulemust (mida saad t-testi tulemuseks ja miks).

2. Milline on Ansari-Bradley teststatistiku jaotus H_0 kehtides, kui $m = 3$ ja $n = 4$?

Peatükk 4

Friedmani test

Vahel on tarvis võrrelda rohkem kui kahte töötlust või populatsiooni. Olgu töötluste mõju katsetatud samadel objektidel. Näiteks kasvatatakse k erinevat viljasorti samal põllul asuvatel erinevatel katselappidel, kokku olgu selliseid katselappideks jagatud põlde n tükki. Või lastakse n hindajal igaühel anda hinnang k -le tootele. Kontrollimaks, kas erinevad töötlusted annavad erinevaid tulemusi või mitte, saab kasutada Friedmani testi. Vahel on Friedmani testi kutsutud ka kahefaktorilise dispersioonanalüüsi mitteparameetriliseks alternatiiviks.

Vaatame järgmist näidet. Kolme sorti vilja kasvatati 4-l põllul. Saadud saagid on toodud alljärgnevas tabelis:

põld	sort 1	sort 2	sort 3
1	989	670	780
2	860	123	870
3	600	345	450
4	520	480	490

Kuidas konstrueerida sobivat mitteparameetrilist testi? Kõigepealt võime igal põllul saadud saagid asendada astakutega — mitmes paremuselt antud sort antud põllul oli:

põld	sort 1	sort 2	sort 3
A	1	3	2
B	2	3	1
C	1	3	2
D	1	3	2

Edaspidistes aruteludes kasutame j . sordi astakut i . põllul tähega R_{ij} .

Leiame iga sordi jaoks keskmise astaku — j . sordi keskmise astaku tähistame sümboliga $\overline{R}_j$:

$$\overline{R}_j := \frac{1}{n} \sum_{i=1}^n R_{ij}$$

Juhul, kui kõik sordid oleksid samahead, peaks astakute jaotuseks olema ühtlane jaotus ning $E\overline{R}_j = \frac{k+1}{2}$. Sestap peaksid vahed $\overline{R}_j - \frac{k+1}{2}$ olema nullhüpoteesi (ühegi töötuse mõju ei erine teiste töötuse mõjust) kehtides nullilähedased. Siis aga peaks nullhüpoteesi kehtides olema suhteliselt väike ka suurus

$$\sum_{j=1}^k \left(\overline{R}_j - \frac{k+1}{2} \right)^2.$$

Friedmani teststatistik, Q , saadakse nimetatud avaldise läbikorrutamisel sobivaid omadusi tagava konstandiga (miks selline konstant on valitud, uurime veidi hiljem):

$$Q = \frac{12n}{k(k+1)} \sum_{i=1}^k \left(\overline{R}_i - \frac{k+1}{2} \right)^2.$$

Toodud näite puhul oleksid $\overline{R}_1 = 5/4 = 1,25$, $\overline{R}_2 = 3$ ja $\overline{R}_3 = 7/4 = 1,75$ ning

$$\begin{aligned} Q &= \frac{12 \cdot 4}{3 \cdot 4} ((1,25 - 2)^2 + (3 - 2)^2 + (1,75 - 2)^2) \\ &= 4 \cdot 1,625 (= 6,5). \end{aligned}$$

4.1 Friedmani teststatistiku jaotus H_0 kehtides

Palju on erinevaid võimalusi astakuid ümber paigutada? Igas grupis (põllul) on $k!$ võimalikku astakute järjestust, kusjuures nullhüpoteesi kehtides (sortidel pole mingit vahet) on kõik nimetatud järjestused võrdtõenäolised. Kuna grupe (põlde) on n tükki, siis on kokku $(k!)^n$ erinevat võimalust paigutada astakuid ringi. Kõik need võimalused on nullhüpoteesi kehtides võrdtõenäolised, st. esinevad tõenäosusega $1/(k!)^n$. Meie näites eksisteerib seega $6^4 = 1296$ erinevat astakute järjestust. Arvuti abil on suhteliselt lihtne kontrollida, et neist selliseid, mille puhul Q väärtus tuleb samasuur kui ülaltoodud näites ($Q = 6,5$) või veel suurem on 54 (6-juhtu, kus kõik sordid saavad kõigil põldudel sama astaku; siis olukord, kus muidu kõik sordid saavad sama astaku, aga ühel põllul on astakud 1 ja 2 vahetusse läinud —

4 põldu $\cdot$ 6 võimalust järjestada astakuid ühel põllul; siis juhtum kus ühel põllul on vahetusse läinud astakud 3 ja 2 — $4 \cdot 6 = 24$ võimalust; kokku $6 + 24 + 24 = 54$ astakute järjestust). Seega toodud näite puhul on Friedmani teststatistiku olulisus $54/1296 = 0,042$ ning meil on võimalik (olulisuse nivool 0,05) kummutada nullhüpoteesi — kõik sordid pole samahead.

4.2 Asümptootiline jaotus nullhüpoteesi kehtides

Friedmani teststatistiku asümptootiline jaotus nullhüpoteesi kehtides on hii-ruut jaotus,

$$Q \xrightarrow{\mathcal{D}} \chi_{df=k-1}^2.$$

Üritame seda näidata. Kõigepealt mainime (ilma tõestuseta) peamisi tulemusi, mida tõesamiseks kasutame.

A. **Tsentraalne piirteoreem** (valimikeskmise jaotus on normaaljaotus). Olgu $\mathbf{X}_i$, $i = 1, \dots, n$ sõltumatud juhuslikud vektorid dimensiooniga k . Kui $E\mathbf{X}_i = \boldsymbol{\mu}$ ja $D\mathbf{X}_i = \boldsymbol{\Sigma}$, siis

$$\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow{\mathcal{D}} N_k(\mathbf{0}, \boldsymbol{\Sigma}).$$

B. Kui $X_n \xrightarrow{\mathcal{D}} X$ ja $f(X) \sim F$, kus $f(x)$ on pidev funktsioon, siis

$$f(X_n) \xrightarrow{\mathcal{D}} F.$$

C. Lisaks vajame tulemust, mis ütleb: Kui $\mathbf{X}$ on juhuslik vektor dimensiooniga k , nii et $\mathbf{X} \sim N_k(\mathbf{0}, \boldsymbol{\Sigma})$ ja $\boldsymbol{\Sigma}$ on idempotentne, st $\boldsymbol{\Sigma}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}$, siis (ja ainult siis) $\mathbf{X}^T \mathbf{X} \sim \chi_{df=\text{rank}(\boldsymbol{\Sigma})}^2$.

Tähistame i . katseüksuse (põld või degustaator) j . töötlustele (väetamine või veinisort) antud astaku R_{ij} . Meenutame kõigepealt, milline on astakute keskväärtus, dispersioon ja kahe erineva (samal põllul mõõdetud) astaku omavahelist kovariatsiooni nullhüpoteesi kehtides (vaata tulemusi Wilcoxon astaksummatesti käsitlevast osast):

$$\begin{aligned} ER_{ij} &= \frac{k+1}{2} \\ DR_{ij} &= \frac{k^2-1}{12} \\ \text{Cov}(R_{ij}, R_{ij'}) &= -\frac{k+1}{12} \end{aligned}$$

Tsentraalsest piirteoreemist järeldub, et

$$\mathbf{U}_n = \sqrt{n} \left(\overline{R}_1 - \frac{k+1}{2}, \dots, \overline{R}_k - \frac{k+1}{2} \right)^T \xrightarrow{\mathcal{D}} N_k(\mathbf{0}, \mathbf{\Sigma}),$$

kus

$$\mathbf{\Sigma} = \begin{pmatrix} \frac{k^2-1}{12} & \cdots & -\frac{k+1}{12} \\ \vdots & \ddots & \vdots \\ -\frac{k+1}{12} & \cdots & \frac{k^2-1}{12} \end{pmatrix}.$$

Paraku pole saadud kovariatsioonimaatriks idempotentne (mida on lihtne kontrollida). Proovime olukorda parandada – korrutades kõik astakud läbi konstandiga $\sqrt{\frac{12}{k(k+1)}}$, saame teisendatud vektori $\sqrt{\frac{12}{k(k+1)}} \mathbf{U}_n$, mille dispersioonimaatriksiks on $\mathbf{\Sigma}^* := \mathbf{D} \sqrt{\frac{12}{k(k+1)}} \mathbf{U}_n = \frac{12}{k(k+1)} \mathbf{D} \mathbf{U}_n$ väärtustega

$$\mathbf{\Sigma}^* = \begin{pmatrix} \frac{k-1}{k} & \cdots & -\frac{1}{k} \\ \vdots & \ddots & \vdots \\ -\frac{1}{k} & \cdots & \frac{k-1}{k} \end{pmatrix}.$$

On lihtne näha, et mainitud maatriks on idempotentne (korruta ta iseendaga läbi!). Järelikult

$$\mathbf{X}_n := \sqrt{\frac{12n}{k(k+1)}} \left(\overline{R}_1 - \frac{k+1}{2}, \dots, \overline{R}_k - \frac{k+1}{2} \right)^T \xrightarrow{\mathcal{D}} N_k(\mathbf{0}, \mathbf{\Sigma}^*).$$

Aga tehnilistest abitulemustest B ja C järeldub seega, et $\mathbf{X}_n^T \mathbf{X}_n \xrightarrow{\mathcal{D}} \chi_{\text{rank}(\mathbf{\Sigma}^*)}^2$. $\mathbf{\Sigma}^*$ astakut on aga lihtne leida — idempotentse maatriksi puhul maatriksi astak langeb kokku maatriksi jäljega. Seega $\text{rank}(\mathbf{\Sigma}^*) = \sum_{i=1}^k \frac{k-1}{k} = k-1$. Olemegi näidanud, et H_0 kehtides $Q \xrightarrow{\mathcal{D}} \chi_{k-1}^2$.

4.3 Friedmani test korduvate väärtuste olemasolu korral

Vahel pole vaatlused üheselt järjestatavad, näiteks siis, kui samal põllul annavad kaks või enam erinevat sorti annavad täpselt sama saagi (*ties*). Sellisel juhul on kombeks kirjutada kokkulangevate vaatluste astakuks nn keskmise astak (*midrank*). Seega kui eri sorti viljade saagid ühel põllul oleksid 290, 310, 310, 325, 325, 325, siis vastavad astakud oleksid 1; 2,5; 2,5; 5; 5; 5. Paraku muutub keskmiste astakute kasutamisel astakute summa dispersioon (väheneb). Selleks, et Friedmani teststatistiku jaotust saaks hii-ruut jaotusega jätkuvalt lähendada, peame kasutama veidi modifitseeritud Friedmani statistikut Q^* :

$$\begin{aligned} Q^* &= Q / \text{konst} \\ &= \frac{12n}{k(k+1)} \sum_{j=1}^k \left(\overline{M}_j - \frac{k+1}{2} \right)^2 / \left(1 - \sum_{i=1}^n \sum_{u=1}^{e_i} \frac{d_{iu}^3 - d_{iu}}{n \cdot k(k^2 - 1)} \right), \end{aligned}$$

kus e_i näitab, mitu erinevat väärtust on mõõdetud i . põllul ning suurus d_{iu} näitavad i . põllul mõõdetud u . unikaalse väärtuse kohta, mitmel korral sellist väärtust tuli ette i . põllul. Tähis $\overline{M}_j$ näitab j . töötuse keskastakute aritmeetilist keskmist:

$$\overline{M}_j = \frac{1}{n} \sum_{i=1}^n M_{ij},$$

kus M_{ij} on i . ploki j . töötuse keskastak.

Selgitamaks suuruseid e_i ja d_{iu} , teeme läbi ühe lihtsa näite. Olgu meie andmestik järgmine:

põld	sort 1	sort 2	sort 3	sort 4	sort 5
1	120	670	670	680	720
2	860	860	870	870	870
3	600	345	450	432	123
..	..	..	..	..	..

Siis $e_1 = 4; e_2 = 2; e_3 = 5$ ja $d_{11} = 1; d_{12} = 2; d_{13} = 1; d_{14} = 1; d_{21} = 2; d_{22} = 3; d_{13} = 1; \dots$. Tasub tähele panna, et kui korduseid andmestikus pole, $d_{ij} = 1$, siis $Q^* = Q$.

4.4 Sobitatud juht-kontrolluuring — McNemar'i test

Vaatame lähemalt ühte praktikas sageli ettetulevat erijuhtu-uuringutüüpi. Haruldaste haiguste põhjuste otsimisel küsitatakse esmalt haigeid, mida nad on söönud/teinud (kas nad suitsetavad, kas nad töötavad uraanikaevanduses jne). Siis otsitakse iga haigega sarnane (sama vana, samast soost, sama rikas jne) terve inimene ja küsitakse temalt samu küsimusi. Hiljem soovetakse siis kontrollida, kas riskifaktor (näiteks suitsetamine) võiks olla seotud (põhjustada) haigestumisega. Taolist uuringutüüpi kutsutakse sobitatud juht-kontrolluuringuks (*matched case-control study*). Vaatame järgmist näidet — oletame, et A juhtum-kontroll paari olid sellised, et nii juhtum kui kontroll suitsetasid, B paari olid sellised, et kontroll (terve) ei suitsetanud, aga juhtum (haige) suitsetas jne, vaata ka alljärgnevat tabelit:

		kontroll	
		suitsetas (1)	ei suitsetanud (0)
juhtum	suitsetas (1)	A	B
	ei suitsetanud (0)	C	D

Milline näeks antud olukorras välja Friedmani teststatistik?

Kuna $k = 2$ siis

$$Q^* = 2n ((\bar{R}_{juhtum} - 3/2)^2 + (\bar{R}_{kontroll} - 3/2)^2) / konst.$$

Jätkamiseks kirjutame välja, millised võimalused on astakute jagunemiseks igas juhtum-kontroll paaris:

vaatlused		astakud		selliste paaride
juhtum	kontroll	juhtum	kontroll	arv
suitsetab	suitsetab	1,5	1,5	A
suitsetab	ei suitseta	2	1	B
ei suitseta	suitsetab	1	2	C
ei suitseta	ei suitseta	1,5	1,5	D

Nüüd näeme, et $\bar{R}_{juhtum} = (1,5A + 2B + 1C + 1,5D)/n$ ja $\bar{R}_{kontroll} = (1,5A + 1B + 2C + 1,5D)/n$ ning

$$\begin{aligned} \bar{R}_{juhtum} - 1,5 &= (1,5A + 2B + 1C + 1,5D)/n - 1,5 \\ &= (1,5A + 2B + 1C + 1,5D - 1,5A - 1,5B - 1,5C - 1,5D)/n \\ &= (0,5B - 0,5C)/n, \end{aligned}$$

sest $n = A + B + C + D$. Analoogselt saame, et

$$\overline{R}_{kontroll} - 1,5 = (0,5C - 0,5B)/n.$$

Seega oleme leidnud, et

$$\begin{aligned} Q^* &= 2n \left(\frac{1}{4n^2} (B - C)^2 + \frac{1}{4n^2} (B - C)^2 \right) / konst \\ &= \frac{1}{n} (B - C)^2 / konst \end{aligned}$$

Nüüd on tarvis veel vaid leida konstandi *konst* väärtust. Selleks paneme tähele, et meil iga sõltuvate vaatluste paari korral on suurustel d_{iu} kaks võimalikku väärtust. Kas $d_{i1} = 2$ ($A + D$ juhtu) või $d_{i1} = 1$ ja $d_{i2} = 1$ ($B + C$ juhtu). Seega

$$\begin{aligned} konst &= \left(1 - \sum_{i=1}^n \sum_{u=1}^{e_i} (d_{iu}^3 - d_{iu}) / (nk(k^2 - 1)) \right) \\ &= 1 - (A + D)(8 - 2) / (n \cdot 2 \cdot (4 - 1)) - (B + C)(1 - 1) / (6n) \\ &= 1 - (A + D)/n \\ &= (B + C)/n. \end{aligned}$$

Lõpptulemuseks seega:

$$\begin{aligned} Q^* &= \frac{1}{n} (B - C)^2 / ((B + C)/n) \\ &= (B - C)^2 / (B + C). \end{aligned}$$

Leitud teststatistik peaks siis nullhüpoteesi kehtides olema (suure n korral) ligikaudu hii-ruut jaotusega,

$$(B - C)^2 / (B + C) \sim \chi_{df=1}^2.$$

Friedmani testi erijuhtu, kus on tegemist kahe grupi võrdlusega (juhud/kontrollid) ning kirjeldavalt tunnusel on vaid kaks erinevat võimalikku väärtust (suitsetas/ei suitsetanud) tuntakse McNemari testi nime all.

Näide

Arsti nimekirjas on 85 Hodgkini tõve (lümfisüsteemi vähkkasvaja) all kannatavat patsienti. Teame, kellel neist on mandlid ära lõigatud ja kellel mitte.

Neil on 85 enam-vähem samavana õve (õde või vend). Ka nende kohta teame, kas neil on mandlid opereeritud või mitte. Andmetabel näeb välja järgmine:

		terve õde või vend	
		Mandlitega	Mandliteta
Haige	Mandlitega	37	7
	Mandliteta	15	26

Kust saame statistiku väärtuse $Q = (15 - 7)^2 / (15 + 7) = 2,909\dots$ ja testi p-väärtuseks tuleb 0,088.

4.5 Post-hoc testid

Juhul, kui Friedmani testi abil õnnestub tõestada alternatiivne hüpotees, siis tekib küsimus — millised sordid siis ikkagi on teistest paremad? Võrdlemaks sorte omavahel (peale alternatiivse hüpoteesi vastuvõtmist) on välja pakutud paar erinevat meetodit. Kõigepealt võime paarikaupa välja valida töötluseid (sorte) ja teha läbi $k \cdot (k - 1)$ võrdlust kasutades Friedmani testi — kuna meetodite paarikaupa võrdlemine Friedmani testi abil on sama mis märgitesti, siis võime kasutada ka märgitesti. Iga testi puhul peaksime siis muidugi kasutama olulisuse nivood $\frac{2\alpha}{k \cdot (k-1)}$, ehk kasutama Bonferroni korrektsiooni.

Suure valimi korral (n on suur) võime kasutada ka järgmist meetodit, vt Daniel (1990). Loe töötluste (sortide) i ja j vaheline erinevus statistiliselt oluliseks, kui

$$|\overline{R}_i - \overline{R}_j| \geq z_\gamma \sqrt{\frac{k(k+1)}{6n}},$$

kus z_γ on standardse normaaljaotuse $\gamma = 1 - \frac{\alpha}{k(k-1)}$ -kvantiil.

Friedmani testi illustreerivas näites peaksime siis uurima, kas keskmiste astakute erinevus tuleb suurem kui $z_{1-0,0083} \sqrt{\frac{3 \cdot 4}{6 \cdot 4}} = 1,693\dots$. Kuna toodud näites on maksimaalne keskmiste astakute erinevus esimese ja teise sordi vahel $|1,25 - 3| = 1,75$, need sordid võime erinevaks lugeda. Teised kaks võrdlust jääks nullhüpoteesi juurde, kuna erinevused $|1,27 - 1,75| = 0,5$ ja $|3 - 1,75| = 1,25$ on alla kriitilist piiri 1,693.

4.6 Ajalooline vahepala

Friedman'i testi autor on tõepoolest kuulsast statistiku Harold Hotelling'i (Hotellingi T^2) õpilane, Milton Friedman. Esimest korda kasutas Friedmani omanimelist testi uurides, kas vaesamad pered kulutavad alati, kõigis valdkondades ühtlaselt vähem kui rikkamad inimesed (või äkki esineb kuluallikaid, mille peale vaesed kulutavad rikkastest rohkem), vaata ka alltoodud andmetabelit tema artiklist:

TABLE I
STANDARD DEVIATIONS AT DIFFERENT INCOME LEVELS* OF EXPENDITURES ON
THE MAJOR CATEGORIES DURING 1935-36 OF 246 MINNEAPOLIS AND
ST. PAUL FAMILIES OF WAGE-EARNERS AND LOWER
SALARIED CLERICAL WORKERS†

Category of expenditure	Annual family income						
	\$750- 1,000	\$1,000- 1,250	\$1,250- 1,500	\$1,500- 1,750	\$1,750- 2,000	\$2,000- 2,250	\$2,250 2,500
Housing	\$103.3	\$68.42	\$89.53	\$77.94	\$100.0	\$108.2	\$184.9
Household operation	42.19	44.31	60.91	73.90	43.87	61.74	102.3
Food	71.27	81.88	100.71	86.52	100.3	90.75	100.6
Clothing	37.59	60.05	56.97	60.79	71.82	83.04	117.1
Furnishings and equipment	58.31	52.73	96.04	60.42	104.33	89.78	85.77
Transportation	46.27	82.18	129.8	181.0	172.33	164.8	246.8
Recreation	19.00	23.07	38.70	45.81	59.03	50.69	55.18
Personal care	8.31	8.43	9.16	14.28	10.63	15.84	12.50
Medical care	20.15	33.48	60.08	69.35	114.34	45.28	101.6
Education	3.16	4.12	12.73	18.95	8.89	41.52	66.33
Community welfare	4.12	18.87	8.54	12.92	25.30	19.85	16.76
Vocation	7.68	11.18	10.44	10.95	10.54	13.96	14.39
Gifts	5.29	10.91	11.22	25.26	42.25	48.80	69.38
Other	6.00	5.57	22.23	2.45	6.24	1.00	4.00

[...]

The sampling distribution of the mean ranks (where $\bar{r}_j$ is the mean rank of the j -th column) will have a mean value (ρ) of $\frac{1}{2}(p+1)$ and a variance σ^2 of $(p^2-1)/(12n)$, where n is the number of rows, i.e., the number of ranks averaged.⁷

Since the true mean and true standard deviation of the chance universe are known, the hypothesis that the means come from a single homogeneous normal universe can be tested by computing

$$\chi_r^2 = \frac{p-1}{p\sigma^2} \sum_{j=1}^p (\bar{r}_j - \rho)^2 = \frac{12n}{p(p+1)} \sum_{j=1}^p \left\{ \bar{r}_j - \frac{1}{2}(p+1) \right\}^2.$$

TABLE II
RANKING OF INCOME LEVELS BY SIZE OF STANDARD DEVIATION FOR EACH
CATEGORY OF EXPENDITURE*

Category of expenditure	Annual family income						
	\$750- 1,000	\$1,000- 1,250	\$1,250- 1,500	\$1,500- 1,750	\$1,750- 2,000	\$2,000- 2,250	\$2,250- 2,500
Housing	5	1	3	2	4	6	7
Household operation	1	3	4	6	2	5	7
Food	1	2	7	3	5	4	6
Clothing	1	3	2	4	5	6	7
Furnishings and equip- ment	2	1	6	3	7	5	4
Transportation	1	2	3	6	5	4	7
Recreation	1	2	3	4	7	5	6
Personal care	1	2	3	6	4	7	5
Medical care	1	2	4	5	7	3	6
Education	1	2	4	5	3	6	7
Community welfare	1	5	2	3	7	6	4
Vocation	1	5	2	4	3	6	7
Gifts	1	2	3	4	5	6	7
Other	5	4	7	2	6	1	3
a. Total	23	36	53	57	70	70	83
b. Mean rank	1.643	2.571	3.786	4.071	5.000	5.000	5.929
c. Deviation	-2.357	-1.429	-.214	.071	1.000	1.000	1.929

Sum of squared deviations = 13.3692

$\chi^2 = 40.108$

[...]

The number of degrees of freedom on which this estimate is based is $p - 1 = 6$. For six degrees of freedom the value of χ^2 which would be exceeded by chance once in 20 times is 12.592, and once in a hundred times, 16.812.⁹ The probability of a value greater than 40 is .000001.¹⁰ There can thus be little question that the observed mean ranks differ significantly, i.e., that the standard deviation is related to the income level. From the mean ranks it is seen that with but one minor exception the standard deviations consistently increase with income.

[...]

Peatükk 5

Kruskal-Wallise test

Kruskal-Wallise test on Wilcoxon astaksummatesti üldistus enam kui kahe valimi jaoks.

Oletame, et meil on k sõltumatut valimit, igaühes neist on tehtud n_i mõõtmist, kokku N vaatlust. Näiteks võrdleme omavahel kolme erinevat tee-sorti ja huvi pakub millise sordi eelistajad kõige kiiremini kõhast lahti saavad. Katsesse kaasatakse 8 inimest: 3 musta tee joojat, 2 roheline tee eelistajat ning 3 spetsiifilise kohate tarbijat. Terveks saamiseni kulunud aeg tundides on katses osalejatel ära toodud järgmises tabelis:

tee	
must	39 45 60
roheline	47 53
kohatee	36 50 61

Gruppide võrdlemiseks paigutame vaatlused taas ühisesse variatsioonritta (nagu Wilcoxon astaksumma testi korral) ja nõnda saab i . valimist pärit j . vaatlus endale järjekorranumbri ehk astaku R_{ij} :

var. rida	36	39	45	47	50	53	60	61
grupp	K	M	M	R	K	R	M	K
tähis	R_{31}	R_{11}	R_{12}	R_{21}	R_{32}	R_{22}	R_{13}	R_{33}
astak	1	2	3	4	5	6	7	8

Iga valimi jaoks leiame tema vaatluste astakute keskmise

$$\overline{R_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} R_{ij}.$$

Kui paigutame leitud astakud gruppide kaupa tabelisse nii, et grupid määravad tabeli read, on vaja leida astakute keskmised tabeli ridades:

tee				$\bar{R}_i$
must	2	3	7	12/3
roheline	4	6		10/2
kõhatee	1	5	8	14/3

Ühe suvaliselt valitud vaatluse keskmine astak on $1/N \sum_{i=1}^N i = (N+1)/2$ ja nullhüpoteesi kehtides (kõik töötlused on võrdsed) on i . grupi oodatavaks keskmiseks astakuks $(N+1)/2$. Teststatistiku leidmiseks liidame iga vaatluse korral tema grupi astakute aritmeetilise keskmise ja oodatava keskmise $(N+1)/2$ erinevuste ruudud ja korrutame saadud summa konstandiga $12/(N(N+1))$. Kruskal-Wallise teststatistikuks on seega

$$K = \frac{12}{N(N+1)} \sum_{i=1}^k n_i \left(\bar{R}_i - \frac{N+1}{2} \right)^2.$$

Meie näite korral

$$\begin{aligned} K &= \frac{12}{8 \cdot 9} \left(3 \left(\frac{12}{3} - 4,5 \right)^2 + 2 \left(\frac{10}{2} - 4,5 \right)^2 + 3 \left(\frac{14}{3} - 4,5 \right)^2 \right) \\ &= \frac{1 \cdot 48}{6 \cdot 36} = \frac{2}{9} \approx 0,222 \end{aligned}$$

Statistiku K eeltoodud valemit teisendades saab leida ka alternatiivse arvutusvalemi, mis on veidi mugavam kasutada kuna ei pea leidma erinevusi oodatatavast keskmisest:

$$K = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1),$$

kus

$$R_i = \sum_{j=1}^{n_i} R_{ij}.$$

5.1 Kruskal-Wallise statistiku jaotus H_0 kehtides

Nullhüpoteesi kehtides peaks $K > 0$ väärtus olema väike, nullilähedane. Alternatiivse hüpoteesi kehtides (kui mõnest populatsioonist pärit vaatlused

kipuvad olema teistest suuremad) aga peaks K väärtus tulema suur (sest erinevus tegeliku keskmise astaku ja H_0 puhul oodatava puhul läheb suureks).

Kuidas leida Kruskal-Wallise testi olulisustõenäosust? Esmalt peaksime leidma, mitu erinevat võimalust on N astakut jagada gruppidesse suurusega $n_1, n_2, \dots, n_k$. Selleks on muidugi

$$\binom{N}{n_1} \cdot \binom{N-n_1}{n_2} \cdot \dots \cdot \binom{N-n_1-\dots-n_{k-1}}{n_k}$$

erinevat võimalust. Kõik need võimalused astakute jagunemiseks k grupi vahel on H_0 kehtides võrdtõenäosed. Seejärel tuleks kokku lugeda, kui paljud neist võimalustest on sellised, mille puhul statistiku K väärtus tuleb suurem (või samasuur) kui meie valimis nähtud K väärtus. Jagades sobivate võimaluste arvu kõigi võimaluste arvuga saamegi teada, millise tõenäosusega H_0 kehtides võiks näha sama ekstreemset või veelgi ekstreemsemat statistiku väärtust ehk oleme leidnud testi olulisustõenäosuse.

Teesortide võrdlemise katse korral on kõiki võimalusi astakute paigutamiseks

$$C_8^3 \cdot C_5^2 \cdot C_3^3 = 560.$$

Nullhüpoteesi kehtides on kõik need võrdvõimalikud. Teststatistiku jaotustabeli algus näeks antud näites välja järgmine:

K	$\frac{1}{36}$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{4}$	$\frac{7}{36}$	$\dots$
sagedus	18	12	28	18	32	$\dots$
tõenäosus	18/560	12/560	28/560	18/560	32/560	$\dots$

Testi olulisuse tõenäosus on tõenäosus, et nullhüpoteesi kehtimise korral saame statistiku väärtuse vähemalt sama ekstreemse kui antud juhul:

$$P\left(K \geq \frac{2}{9}\right) = 1 - P\left(K < \frac{2}{9}\right) = 1 - \frac{30}{560} = \frac{53}{56} \approx 0,946.$$

Seega erinevatest teesortidest saadud abi kõha vastu pole oluliselt erinev.

5.2 Statistiku asümptootiline jaotus H_0 kehtides

Teoorias on kõik ilus, aga praktikas läheb erinevate astakute konfiguratsioonide arv ruttu astronoomiliseks. Kui näiteks meie valimi suurused on $n_1 = 10, n_2 = 11, n_3 = 8$, siis võimalikke viise astakuid jagada neisse kolme gruppi oleks

$$\binom{29}{10} \cdot \binom{19}{11} \cdot 1 = 20030010 \cdot 75582 \approx 1,5 \cdot 10^{12}.$$

Sedavõrd paljusid võimalusi läbi vaadata on isegi arvutilgi raske. Mis siis veel saab, kui valimimahud hakkavad muutuma "mõistlikult" suureks?

Lahenduseks on muidugi asümptootilise jaotuse kasutamine. Selgub, et asümptootiliselt (ja nullhüpoteesi kehtides) on K jaotuseks $\chi^2_{df=k-1}$.

Proovime seda tulemust ka pool-tunnetuslikult tõestada. Tähistame $\mathbf{X}$ -ga järgmise vektori:

$$\mathbf{X} := \left(\sqrt{\frac{12n_1}{N(N+1)}}(\bar{R}_1 - (N+1)/2), \dots, \sqrt{\frac{12n_k}{N(N+1)}}(\bar{R}_k - (N+1)/2) \right)$$

ja paneme tähele, et $K = \mathbf{X}^T \mathbf{X}$. Kui juhusliku vektori $\mathbf{X}$ jaotuseks oleks normaaljaotus, $\mathbf{X} \sim N_k(\mathbf{0}, \mathbf{\Sigma})$, siis vektori $\mathbf{X}^T \mathbf{X}$ jaotuseks on χ^2 -jaotus (siis ja ainult) siis, kui $\mathbf{\Sigma}$ on idempotentne, $\mathbf{\Sigma}\mathbf{\Sigma} = \mathbf{\Sigma}$. Üritame aru saada, milline on meie poolt defineeritud $\mathbf{X}$ -i jaotus. Kuna $\sqrt{n}(\bar{X}_i - \mu_i)$ jaotus on tsentraalse piirteoreemi tõttu normaaljaotus, siis võime oletada, et meie poolt defineeritud vektori kõigi elementide asümptootiliseks jaotuseks on normaaljaotus. Saab näidata, et kogu vektor on mitmemõõtmelise normaaljaotusega (jätame tõestuse hetkel vahele). See, et $\mathbf{X}$ keskvärtus on (nullhüpoteesi kehtides) $\mathbf{0}$, on lihtne näha. Aga milline on $\mathbf{\Sigma}$ ja kas ta on ka idempotentne? Kõigepealt vaatame millised näevad välja $\mathbf{\Sigma}$ diagonaali elemendid ehk milline on DX_i .

$$\begin{aligned} DX_i &= D \left(\sqrt{\frac{12n_i}{N(N+1)}}(\bar{R}_i - (N+1)/2) \right) \\ &= \frac{12n_i}{N(N+1)} D \left(\frac{1}{n_i} \sum_{j=1}^{n_i} R_{ij} \right) \\ &= \frac{12n_i}{N(N+1)n_i^2} \left\{ \sum_{j=1}^{n_i} D(R_{ij}) + \sum_{j \neq j'} Cov(R_{ij}, R_{ij'}) \right\} \\ &= \frac{12}{N(N+1)n_i} \left\{ n_i \frac{N^2 - 1}{12} + n_i(n_i - 1) \left(-\frac{N+1}{12} \right) \right\} \\ &= \frac{12}{N(N+1)n_i} \left\{ \frac{(N+1)n_i}{12} (N - 1 - (n_i - 1)) \right\} \\ &= \frac{N - n_i}{N} \end{aligned}$$

ja väljaspool maatriksi peadiagonaali asuv element $\Sigma_{ij} = Cov(X_i, X_j)$ näeb

välja selline:

$$\begin{aligned}
 \text{Cov}(X_i, X_j) &= \sqrt{n_i n_j} \frac{12}{N(N+1)} \text{Cov}(\overline{R_i}, \overline{R_j}) \\
 &= \sqrt{n_i n_j} \frac{12}{N(N+1)} \frac{1}{n_i n_j} \sum_{k=1 \dots n_i, l=1 \dots n_j} \text{Cov}(R_{ik}, R_{jl}) \\
 &= \sqrt{n_i n_j} \frac{12}{N(N+1)} \left(-\frac{N+1}{12} \right) \\
 &= -\frac{\sqrt{n_i n_j}}{N}
 \end{aligned}$$

Seega on vektori $\mathbf{X}$ dispersioonimaatriks Σ kujul

$$\Sigma = \begin{pmatrix} \frac{N-n_1}{N} & \dots & -\frac{\sqrt{n_1 n_k}}{N} \\ \vdots & \ddots & \vdots \\ -\frac{\sqrt{n_k n_1}}{N} & \dots & \frac{N-n_k}{N} \end{pmatrix}.$$

Siit on juba näha, et Σ on idempotentne — Σ^2 diagonaali i element on

$$\begin{aligned}
 \Sigma_{ii}^2 &= \sum_{j=1}^k \Sigma_{ij} \Sigma_{ji} \\
 &= (\text{D}X_i)^2 + \sum_{j \neq i} \text{Cov}(X_i, X_j)^2 \\
 &= \frac{(N-n_i)^2}{N^2} + \sum_{j \neq i} \frac{n_i n_j}{N^2} \\
 &= \frac{(N-n_i)^2 + n_i(N-n_i)}{N^2} \\
 &= \frac{(N-n_i)(N-n_i+n_i)}{N^2} \\
 &= \frac{N-n_i}{N}
 \end{aligned}$$

ja väljaspool diagonaali paiknev element Σ_{ij}^2 näeb välja ju selline:

$$\begin{aligned}
 \Sigma_{ij}^2 &= \sum_{l=1}^k \Sigma_{il} \Sigma_{lj} \\
 &= (DX_i) \text{Cov}(X_i, X_j) + (DX_j) \text{Cov}(X_i, X_j) + \sum_{l \neq i, j} \text{Cov}(X_i, X_l) \text{Cov}(X_l, X_j) \\
 &= \frac{N - n_i}{N} \left(-\frac{\sqrt{n_i n_j}}{N} \right) + \frac{N - n_j}{N} \left(-\frac{\sqrt{n_i n_j}}{N} \right) + \sum_{l \neq i, j} -\frac{\sqrt{n_i n_l}}{N} \left(-\frac{\sqrt{n_l n_j}}{N} \right) \\
 &= -\frac{(N - n_i + N - n_j) \sqrt{n_i n_j}}{N^2} + \frac{\sqrt{n_i n_j} (N - n_i - n_j)}{N^2} \\
 &= \frac{\sqrt{n_i n_j} (N - n_i - n_j - (N - n_i + N - n_j))}{N^2} \\
 &= -\frac{\sqrt{n_i n_j}}{N}.
 \end{aligned}$$

Seega on Σ idempotentne ja $\mathbf{X}^T \mathbf{X}$ on χ^2 -jaotusega vabadusastmete arvuga $df = \text{rank}(\Sigma) = k - 1$.

5.3 Veel mõningaid tulemusi

Võrdsete vaatluste olemasolu korral võib kasutada keskmiseid astakuid (*mid-rank*), kuid lisaks tuleb jagada statistiku väärtus hajuvuse vähenemist kirjeldava kordajaga:

$$K^* = K / \text{const},$$

kus

$$\text{const} = 1 - \sum_{u=1}^e (d_u^3 - d_u) / (N^3 - N).$$

Toodud valemis võetakse summa üle kõigi unikaalsete väärtuste, mida on kokku e , ja d_u näitab, mitu korda esines u . unikaalne vaatlus valimis.

Post-hoc testid. Kui Kruskal-Wallise testi põhjal on vastu võetud sisukas hüpoteesi, siis otsustamaks millised grupid omavahel erinevad peaks läbi vaatama kõikvõimalikud võrdlusgruppide paarid. Sel juhul on tegu kahe grupi võrdlemistega, kindlasti peaks iga üksiku testi juures kasutama korregeeritud olulisuse nivood. Suure valimi korral võib kasutada ka järgmist meetodit: loe kaks töötlust/populatsiooni teineteisest erinevaks, kui

$$|\overline{R}_i - \overline{R}_l| \geq z_{1 - \frac{\alpha}{k(k-1)}} \sqrt{\frac{N(N+1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_l} \right)}.$$

Peatükk 6

Tihedusfunktsiooni ja jaotusfunktsiooni hindamine

6.1 Jaotusfunktsiooni hindamine

Jaotusfunktsiooni saab hinnata kasutades empiirilist jaotusfunktsiooni,

$$\begin{aligned}\hat{F}(x) &= \#\{X_i \leq x\}/n \\ &= 1/n \cdot \sum_{i=1}^n I(X_i \leq x),\end{aligned}$$

kus I tähistab indikaatorfunktsiooni.

Valimi 1, 4, 6, 9 jaoks leitud parameetriline (leitud normaaljaotuse eeldust kasutades) ja mitteparameetriline jaotusfunktsiooni hinnang on toodud joonisel 6.1.

Empiiriline jaotusfunktsioon on suurepärase omadustega. Tegemist on nihketa hinnanguga (sest suhteline sagedus on nihketa hinnang tõenäosusele):

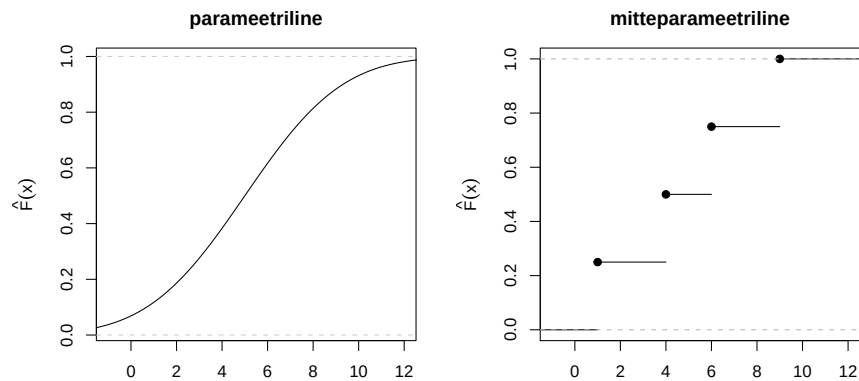
$$\begin{aligned}E\hat{F}(x) &= E\left(1/n \cdot \sum_{i=1}^n I(X_i \leq x)\right) \\ &= 1/n \cdot \sum_{i=1}^n E(I(X_i \leq x)) \\ &= 1/n \cdot \sum_{i=1}^n F(x) \\ &= F(x)\end{aligned}$$

ja empiiriline jaotusfunktsioon osutub ka mõjusaks hinnanguks (sõltumatute vaatluste korral):

$$\begin{aligned}
 D\hat{F}(x) &= D\left(1/n \cdot \sum_{i=1}^n I(X_i \leq x)\right) \\
 &= 1/n^2 \cdot \sum_{i=1}^n D(I(X_i \leq x)) \\
 &= 1/n^2 \cdot \sum_{i=1}^n F(x)(1 - F(x)) \\
 &= \frac{F(x)(1 - F(x))}{n}.
 \end{aligned}$$

Saadud avaldis koondub aga üsna ilmselt nulliks kui $n \rightarrow \infty$. Järelikult koondub empiiriline jaotusfunktsioon valimi kasvades tegelikuks jaotusfunktsiooniks (tegemist on mõjusa hinnanguga).

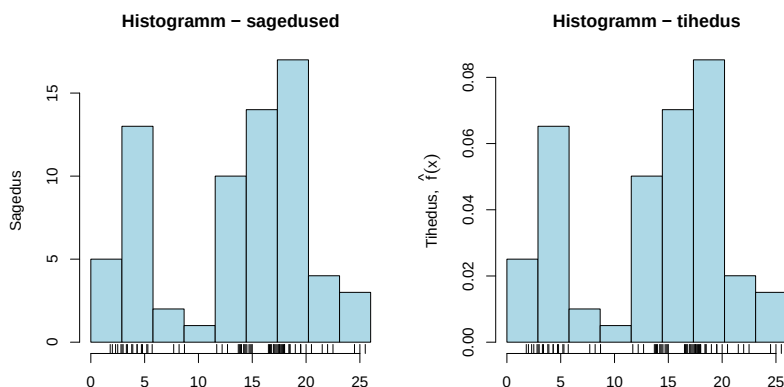
Joonis 6.1: Jaotusfunktsiooni hinnang



6.2 Tihedusfunktsiooni hindamine

Tihedusfunktsiooni mitteparameetrilise hindamisega on mingil kujul arvatavasti iga andmeanalüüsiga tegelenud inimene kokku puutunud — näiteks uurides histogrammi. Tõsi, sageli joonistatakse histogramm küll selliselt, et graafiku y-teljel on mingisse vahemikku sattunud vaatluste arv. Tihedusfunktsiooni hinnangu saamiseks peaksime aga muutma y-telge selliselt, et histogrammi alune pindala oleks üks. Kui kõik histogrammi tulbad on sama laiad, laiusega Δ , ja i . vahemikus on Y_i vaatlust, siis tavalise, sagedusi kasutava histogrammi alune pindala on $\sum_i \Delta \cdot Y_i = \Delta \cdot \sum_i Y_i = \Delta \cdot n$, kus n on vaatluste koguarv. Histogrammist tihedusfunktsiooni hinnangu saamiseks tuleks lihtsalt vaadeldud sagedusi jagada vaatluste arvu ja vahemiku pikkuse korrutisega, vaata ka joonist 6.2.

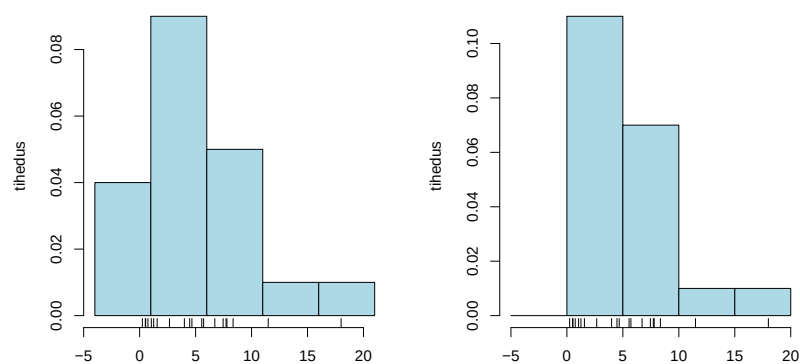
Joonis 6.2: Histogramm - lihtne võimalus hinnata tihedusfunktsiooni



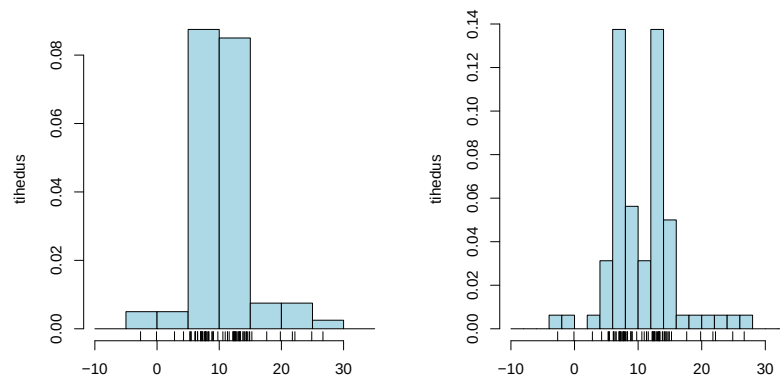
Histogrammi joonistamisel tekivad aga paar rasket küsimust. Kuhu tõmmata tulpade vahelised piirid (vahemike asukohad)? Kui palju vahemikke kasutada (millise laiusega tulpe kasutada)? Vastustest nendele küsimustele võib graafik (ja tema interpretatsioon) sõltuda märkimisväärselt! Vaata ka jooniseid 6.3 ja 6.4.

Lisaks neile kahele küsimusele (vahemike alguspunkti ja laiuse valik) võime muret tunda ka selle üle, kas ehk ei eksisteeri veel mõni teine — loode-tavasti täpsem — võimalus hinnata tihedusfunktsiooni. Vaatamegi, kas oskame välja pakkuda mõnda alternatiivset võimalust tihedusfunktsiooni hindamiseks.

Joonis 6.3: Samad andmed — erinev vahemiku alguspunkt



Joonis 6.4: Samad andmed — erinev tulba laius



6.2.1 Tihedusfunktsiooni hindamine empiirilise jaotusfunktsiooni abil

Jaotusfunktsioonile on olemas heade omadustega (nihketa, mõjus) hinnang — empiiriline jaotusfunktsioon. Kuna jaotusfunktsiooni teades saab leida tihedusfunktsiooni, $F'(x) = f(x)$, siis äkki on võimalik kasutada empiirilist jaotusfunktsiooni tihedusfunktsiooni hindamiseks? Empiirilise jaotusfunktsiooni tuletis meile midagi väga mõistlikku ei anna (milline on treppfunktsiooni tuletis?). Tuletame aga meelde, kuidas numbrilisi meetodeid kasutades saab hinnata/lähendada funktsiooni tuletist — funktsiooni muutumise kiirust — punktis x :

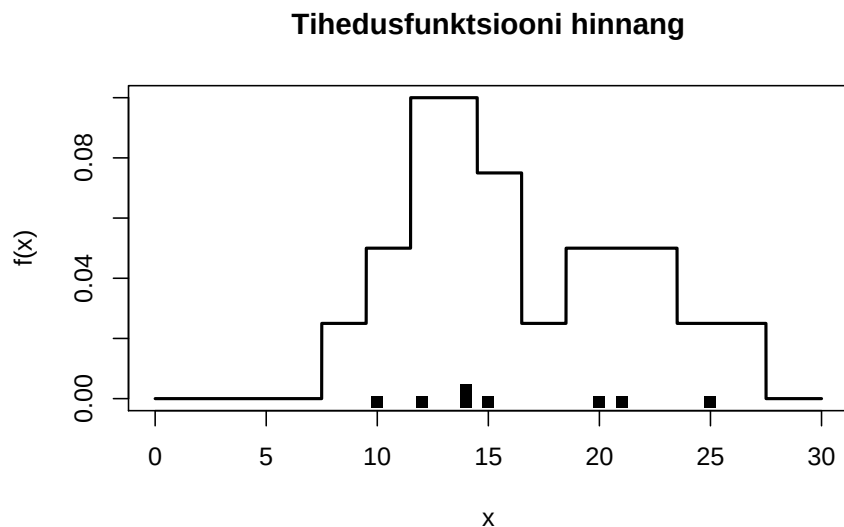
$$F'(x) = \lim_{h \rightarrow 0} \frac{F(x+h) - F(x-h)}{2h}.$$

Siit võiks tulla mõttele kasutada tihedusfunktsiooni hindamiseks kohas x avaldist:

$$\hat{f}(x) = \frac{\hat{F}(x+h) - \hat{F}(x-h)}{2h}.$$

Saadud valemi paremaks mõistmiseks paneme tähele, et $\hat{F}(x+h) - \hat{F}(x-h)$ on vahemikku $(x-h \dots x+h]$ sattunud vaatluste arv jagatud valimi suurusga (ehk antud vahemikku sattunud vaatluste osakaal kõigist vaatlustest). Tihedusfunktsiooni hinnanguks mingis kohas oleks seega seda kohta ümbritsevasse vahemikku sattunud vaatluste arv jagatud valimi suuruse ja vahemiku laiuse korrutisega. Saadud arvutuseeskiri on väga sarnane sellele, milleni jõudsimme histogrammi abil tihedusfunktsiooni hinnates. Kui võrdleksime empiirilist jaotusfunktsiooni kasutatavat hinnangut (kasutades vahemikku laiusega $2h$) ja histogrammi abil saadud hinnangut (histogrammi tulbad laiusega $\Delta = 2h$), siis selgub, et histogrammi tulpade keskpunktides langevad saadud hinnangud alati kokku. Võluv on see, et kokkulangemine toimub alati, ükskõik millise alguspunkti me histogrammi vahemikele ka ei valiks. Seega oleme leidnud meetodi, mille puhul ei kerki enam üles küsimust vahemike alguspunktide paigutamise kohta. Üks lihtne andmestik ja selle andmestiku põhjal empiirilist jaotusfunktsiooni kasutades saadud hinnang tihedusfunktsioonile on toodud pildil 6.5.

Joonis 6.5: Empiirilisest jaotusfunktsiooni kasutatav tihedusfunktsiooni hinnang



Saadud tihedusfunktsiooni hinnangu võime soovi korral kirjutada veidi teisele kujule:

$$\begin{aligned}
 \hat{f}(x) &= \frac{\hat{F}(x+h) - \hat{F}(x-h)}{2h} \\
 &= \frac{\sum_{i=1}^n I(X_i - h < x \leq X_i + h)/n}{2h} \\
 &= \frac{1}{n} \sum_{i=1}^n \frac{I(X_i - h < x \leq X_i + h)}{2h}.
 \end{aligned}$$

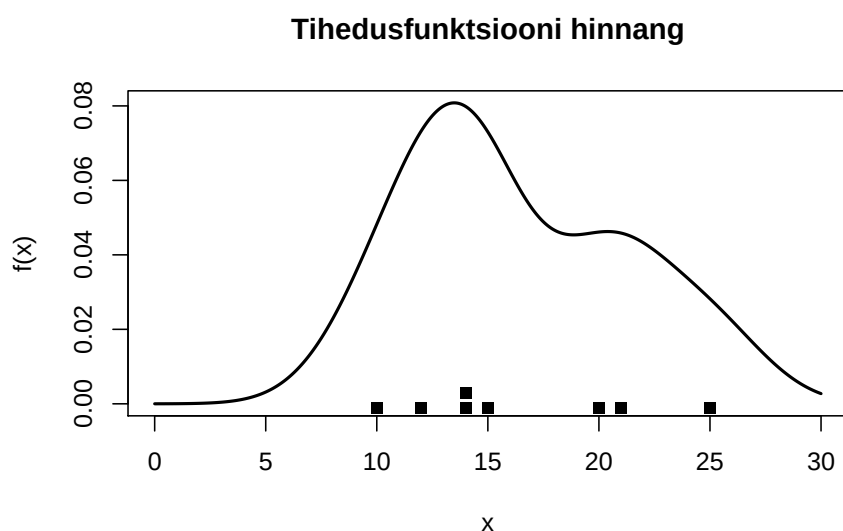
Aga

$$\frac{I(X_i - h < x \leq X_i + h)}{2h}$$

on ühtlase jaotuse $U(X_i - h, X_i + h)$ tihedusfunktsioon, seega on saadud tihedusfunktsiooni hinnang vaadeldav kui n erinevas kohas paikneva ühtlase jaotuse tihedusfunktsiooni keskmine. Tulemuseks on loomulikult tihedusfunktsioon, sest kõik funktsioonid, mille keskmist me leiame, on mittenegatiivsed ja nende alune pindala on 1, seega on ka samade omadustega saadud "keskmine". Samas võiks ju kasutada ühtlase jaotuse tihedusfunktsioonide asemel mõnda teist tihedusfunktsiooni, näiteks võiksime leida tihedusfunktsiooni kui tiheduste $N(x_1, h)$, $N(x_2, h)$ jne keskmise. Juhul, kui kasutame ühtlase

jaotuse asemel vaatluste X_i ümber tsentreeritud normaaljaotuseid, saaksime varem kasutatud andmete (pilt 6.5) põhjal uue hinnangu tihedusfunktsioonile, vaata pilti 6.6. Sellist lähenemist — tihedusfunktsiooni hindamist paljude tihedusfunktsioonide keskmise abil — uurime järgnevas alampeatükis veidi põhjalikumalt.

Joonis 6.6: Tihedusfunktsiooni hinnang kui paljude normaaljaotuste keskmise



6.2.2 Tuumameetod tihedusfunktsiooni hindamiseks

Olgu $K(x)$ mingi keskvärtusega 0 ja lõpliku dispersiooniga juhusliku suuruse tihedusfunktsioon. Vahel nõutakse täiendavalt, et tegemist oleks sümmeetrilise jaotusega, $K(x) = K(-x)$.

Siis funktsiooni

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

kutsutakse tuumameetodi abil leitud tihedusfunktsiooni hinnanguks (*Kernel density estimator, KDE*). Tihedusfunktsiooni hindamisel kasutatud tihedusfunktsiooni $K(x)$ nimetatakse tuumafunktsiooniks (*Kernel function*).

Valemi selgituseks: Olgu antud juhuslik suurus X tihedusega $K(x)$, kesk­väärtusega 0 ja mingi standardhälbega (näiteks $\sigma = 1$). Siis $Y := hX + \mu$ oleks kesk­väärtusega μ ja standardhälbega h (või $h\sigma$, kui $\sigma \neq 1$) juhuslik suurus. Aga juhusliku suuruse Y tihedusfunktsioon $f_Y(y)$ avaldub kujul

$$f_Y(y) = \frac{1}{h} K\left(\frac{y - \mu}{h}\right). \quad (6.1)$$

Selles veendumiseks võid meelde tuletada, et juhusliku suuruse X funktsiooni $Y = g(X)$ tihedusfunktsioon avaldub kujul

$$f_Y(y) = \left| \frac{1}{g'(g^{-1}(y))} \right| \cdot f_X(g^{-1}(y)).$$

Praegusel juhul aga $g(x) = hx + \mu$, $g'(x) = h$ ja $g^{-1}(y) = (y - \mu)/h$. Seega saamegi tulemuseks valemi 6.1.

Kui on antud kesk­väärtusega 0 ja dispersiooniga σ juhuslik suurus X tihedusega $K(x)$, siis $f_Y(x) = \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$ on kesk­väärtusega X_i (ja stan­dardhälbega $h\sigma$) juhusliku suuruse tihedusfunktsioon.

Tuumahinnangu saamiseks tihedusfunktsioonile võtame seega mingi ette­antud hajuvusega $h\sigma$ tihedusfunktsioone, tõstame nad olemasolevate vaat­luste kohale (nii et i . tiheduse kesk­väärtus oleks X_i), leiame saadud tihe­dusfunktsioonide summa ja jagame vaatluste arvuga (et tulemuseks saadud funktsiooni graafiku alune pindala ikka 1 tuleks, nagu tihedusfunktsioonile kohane).

Millise tulemuse saame, jääb muidugi mingil määral sõltuma nii valitud tuumafunktsioonist kui ka kasutatud h väärtusest. Joonisel 6.7 on näha paa­ri enamkasutatavat tuumafunktsiooni, joonisel 6.8 võib aga näha, milliste tihedusfunktsiooni hinnanguteni (sama andmestikku kasutades) jõuame eri­nevate tuumafunktsioonide korral. Sagedamini kasutamist leidvad tuuma­funktsioonid on

- Ühtlane jaotus $U(-1,1)$:

$$K(x) = 0,5I(|x| \leq 1)$$

- Triangular:

$$K(x) = (1 - |x|)^3 I(|x| \leq 1)$$

- Normaalkaotus $N(0,1)$:

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$$

- Epanechnikov'i jaotus:

$$K(x) = \frac{3}{4}(1 - x^2)I(|x| \leq 1)$$

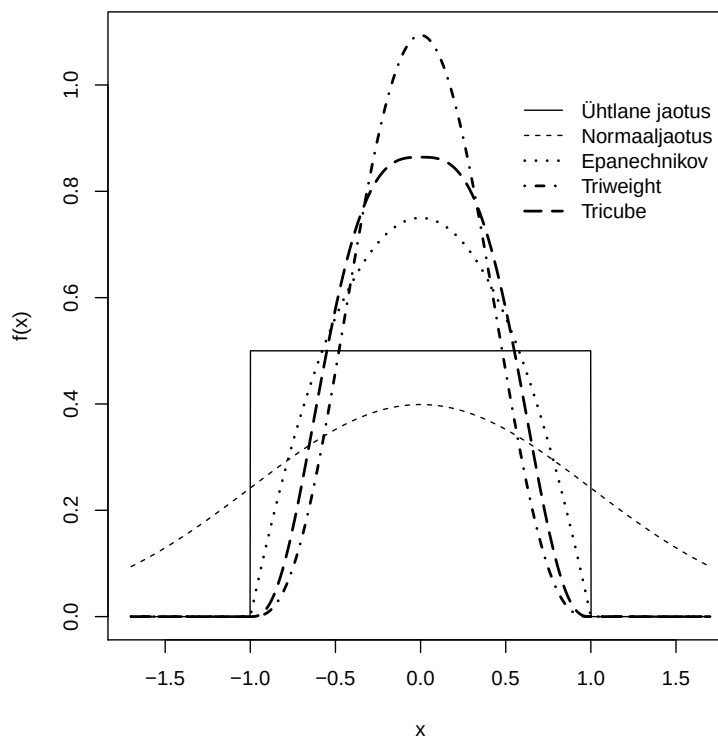
- Triweight:

$$K(x) = \frac{35}{32}(1 - x^2)^3I(|x| \leq 1)$$

- Tricube:

$$K(x) = \frac{70}{81}(1 - |x|^3)^3I(|x| \leq 1).$$

Joonis 6.7: Sagedamini kasutatavaid tuumafunktsioone

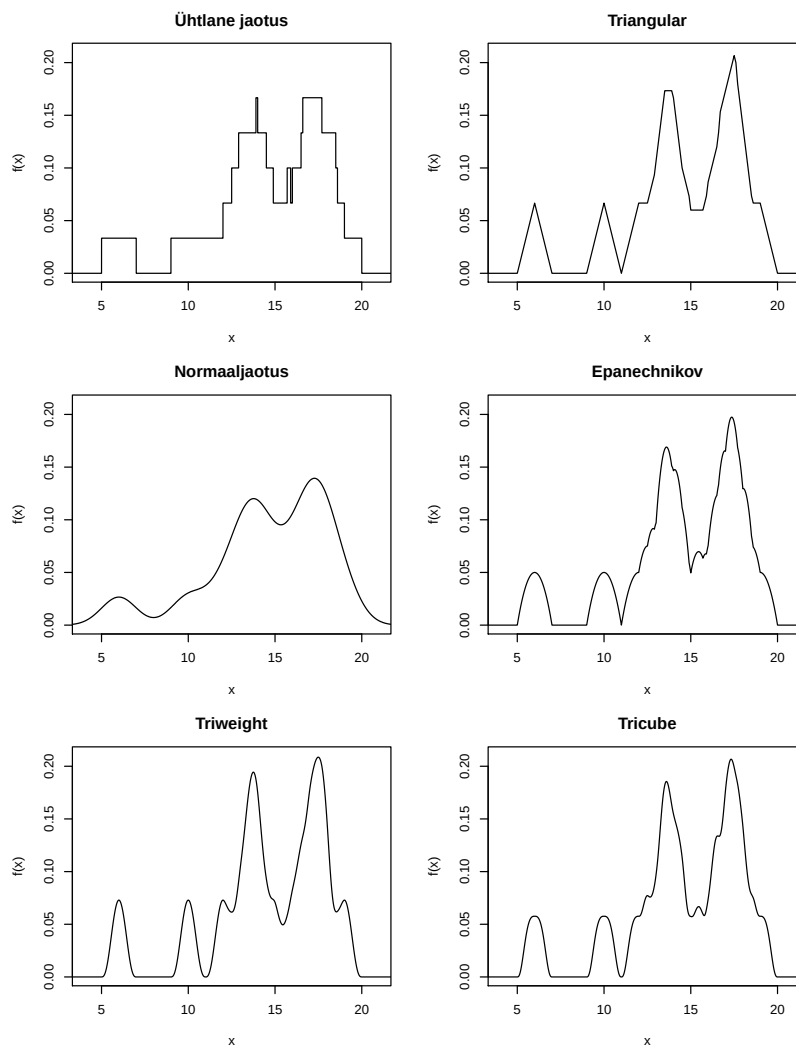


Normaaljaotuse tihedusfunktsioon ei muutu kunagi päris nulliks — see-
ga saadakse normaaljaotust tuumafunktsioonina kasutades selline tihedus-
funktsiooni hinnang, kus ka vägaväga suurte ja väga-väga väikeste väärtuste
saamine on põhimõtteliselt võimalik (kuigi ebatõenäoline). Seevastu Epa-
nechnikovi, triweight, tricube ja ühtlase jaotuse puhul ei mõjuta kaugemal

kui h ühikut paiknevad vaatlused tihedusfunktsiooni hinnangut antud punktis (muul moel kui valimi suuruse kaudu).

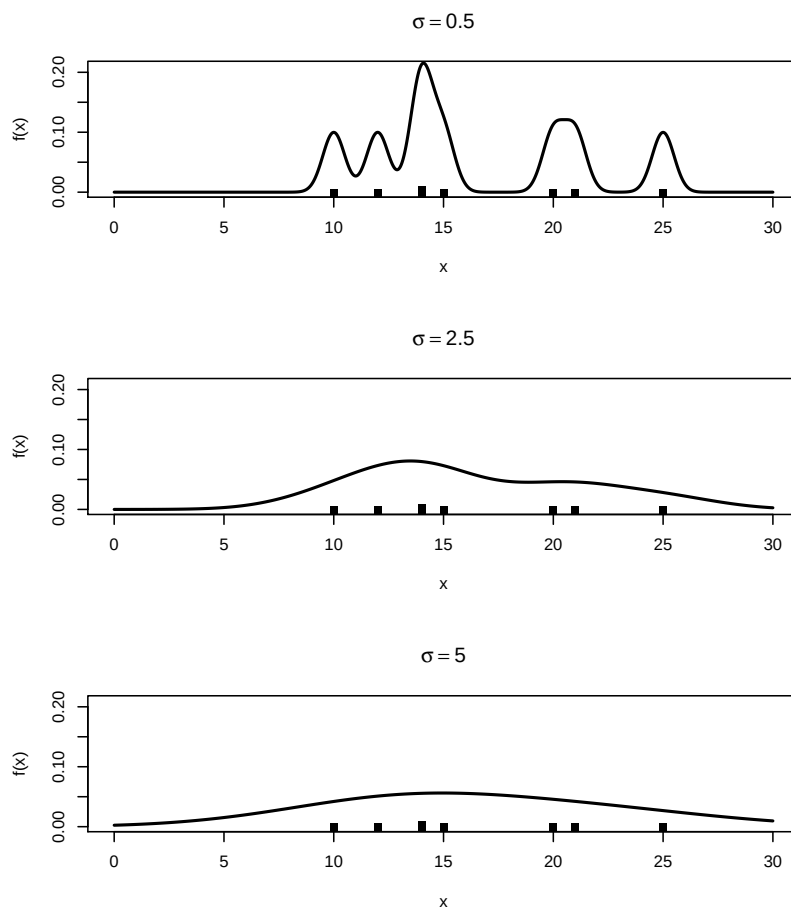
Paraku jääb hetkel veel lahtiseks küsimus, kui suure hajuvusega jaotust — või kui suurt “akna” laiust h — peaksime kasutama (olgu ta siis ühtlane, normaaljaotus või midagi muud). Tulemused võivad tulla üsnagi erinevad, sõltuvalt tehtud valikust. Vaata ka pilti 6.9. Siludes liiga palju — valides

Joonis 6.8: Erinevaid tuumafunktsioone kasutavad tihedusfunktsiooni hinnangud



suure h -i (histogrammi puhul Δ) väärtuse — saame küll väikese dispersiooniga hinnangu (paljude binaarsete — X_i kuulub antud vahemikku või ei — vaatluse keskmine annab üsna täpse hinnangu antud vahemikku sattumise tõenäosusele) aga see-eest võib märkamatuks jääda tegeliku tihedusfunktsiooni mõni tipp või lohk. Liiga väikese h puhul on meil põhimõtteliselt võimalik saada küll kiiremini muutuvat tihedusfunktsiooni hinnangut, mis võiks ju sobituda paremini tegeliku tihedusfunktsiooniga, aga paraku läheb suureks ka valimi juhuslikkusest tingitud hinnanguviga.

Joonis 6.9: Tihedusfunktsiooni hinnang kui paljude normaaljaotuste keskmine



Sobivat akna laiust h võiksime otsida selliselt, et teda kasutades saadud hinnang $\hat{f}(x)$ oleks võimalikult täpne, ehk ruutviga $(\hat{f}(x) - f(x))^2$ oleks võimalikult väike. Paraku on hinnangu viga iga x väärtuse korral erinev, meie sooviksime aga kogu erinevust iseloomustada ühe numbri abil — siis saaksime seda tekkivat “üldist” viga ka minimiseerida. Selleks üheks üldiseks näitajaks võiks sobida näiteks integreeritud ruutviga (*ISE* - *Integrated Squared Error*):

$$ISE = \int (\hat{f}(x) - f(x))^2 dx$$

või keskmine integreeritud ruutviga (MISE), mille puhul minimiseeritakse keskmist ISE't üle kõikmõeldavate valimite:

$$MISE = E \int (\hat{f}(x) - f(x))^2 dx.$$

Mainitud statistikute (ISE ja MISE) arvutamise jäame praktikas muidugi hätta, sest nende leidmiseks peaksime teadma tegelikku tihedusfunktsiooni $f(x)$. Küll aga saab neid näitajaid kasutada teoreetiliste tulemuste leidmiseks (Näiteks: kui suurendame valimi mahtu n , siis MISE väheneb kõige kiiremini, kui valime akna laiuse — ehk tuumafunktsiooni hajuvuse — järgmise eeskirja järgi: $h = cn^{-1/5}$, kus c on mingi hinnatavast tihedusfunktsioonist ja tuumafunktsioonist sõltuv konstant...).

Õnneks selgub siiski, et mõlemad näitajad on (suhteliselt) lihtsasti hinnatavad — konstandi täpsuseni küll. Nimelt:

$$\begin{aligned} ISE &= \int (\hat{f}(x) - f(x))^2 dx \\ &= \int \hat{f}(x)^2 dx - 2 \int \hat{f}(x) \cdot f(x) dx + \int f(x)^2 dx \\ &= \int \hat{f}(x)^2 dx - 2E\hat{f}(x) + \int f(x)^2 dx \\ &\approx \int \hat{f}(x)^2 dx - 2\overline{\hat{f}(x)} + \int f(x)^2 dx. \end{aligned}$$

Kui kaalume, millist aknalaiust h kasutada või millist tuumafunktsiooni $K(x)$ eelistada, siis viimane liidetavatest ei muutu — viimane liidetavatest ei sõltu ju üldse hinnangust $\hat{f}(x)$. Seega võime ta minimiseerimisülesandest eemaldada. Ülejäänud on aga kõik arvutatav — muidugi kui asendame kesk-väärtuse valimi keskmisega, $E\hat{f}(x) \approx \overline{\hat{f}(x)}$.

Tõsi küll, nihketa hinnangu saamiseks $E\hat{f}(x)$ -le tuleks valimi keskmine seekord leida veidi ebatavalisel teel. Nimelt tähistame $\hat{f}_{-i}$ -ga tihedusfunktsiooni hinnangu, mille saame i -ndat vaatlust kasutamata (eemaldame valimist i . vaatluse ja leiame siis tihedusfunktsiooni hinnangu). Siis $\overline{\hat{f}(x)} =$

$\frac{1}{n} \sum_{i=1}^n \hat{f}_{-i}(X_i)$ on nihketa hinnanguks $E\hat{f}(x)$ -le. Saadud meetodit — vali h vastavalt eeskirjale

$$h = \arg \min \int \hat{f}(x)^2 dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{-i}(X_i)$$

tuntakse nihketa ristvalideerimise nime all (*unbiased cross-validation*).

6.2.3 Hinnangu teoreetilised omadused

Järgnevalt toome ära mõned teoreetilised tulemused, mis aitavad välja tuua tuumameetodi eeliseid.

Oletame, et soovime minimiseerida keskmist integreeritud ruutviga (MISE). Siis saab teatud matemaatilistel eeldustel (tihedusfunktsiooni tuletis on pidev jt) näidata, et histogrammi tulba optimaalne laius (tulba laius mille puhul keskmine integreeritud ruutviga üle võimalike valimite tuleb kõige väiksem) on kujul $h_{opt} = n^{-1/3} \cdot c_1$. Kui histogrammi tegemisel kasutada optimaalset tulba laiust, siis $MISE \sim c_2 n^{-2/3}$. Kordajad c_1 ja c_2 on teatavad konstandid, mis sõltuvad hinnatavast tihedusfunktsioonist endast.

Kui aga kasutada tuumameetodit tihedusfunktsiooni hindamiseks, siis $h_{opt} = n^{-1/5} \cdot c_3$ ja optimaalset “venitajat” kasutatades saavutatakse $MISE \sim c_4 n^{-4/5}$, kus c_3 ja c_4 on tuumafunktsioonist ja tegelikust tihedusfunktsioonist sõltuvad konstandid (taas on kasutatud teatud matemaatilisi eelduseid, näiteks peab tegeliku tihedusfunktsiooni teine tuletis olema pidev jne).

Miks mainitud teoreetilised tulemused on tähtsad? Kuna $n^{-4/5}$ koondub nulliks kiiremini kui $n^{-2/3}$, siis vähemalt mingist valimi suurusest alates on tuumafunktsioonil baseeruv hinnang täpsem kui histogrammil baseeruv tihedusfunktsiooni hinnang.

Stone'i teoreem väidab, et asümptootiliselt on ristvalideerimise abil saadud aknalaius sama hea (viib sama täpsete hinnanguteni ISE-mõttes) kui optimaalne aknalaius:

Teoreem 6.1 (Stone) Olgu f tõkestatud. Olgu $\hat{f}_h$ aknalaiust h kasutades saadud hinnang tihedusfunktsioonile. Tähistame ristvalideerimise abil saadud optimaalset aknalaiust h_* -ga. Siis

$$\frac{\int (f(x) - \hat{f}_{h_*})^2 dx}{\inf_h \int (f(x) - \hat{f}_h)^2 dx} \xrightarrow{a.s.} 1.$$

Lisaks saab tõestada tulemuse, et kui informatsiooni tegeliku tihedusfunktsiooni kohta tõepoolest napib, siis pole võimalik leida hinnangut tihedusfunktsioonile, mille MISE (keskmine integreeritud ruutviga) koonduks nulliks kiiremini kui $n^{-4/5}$ (see ei tähenda otseselt, et tuumameetodil saadud hinnang oleks täpsem võimalik, aga see tähendab, et ühelgi hinnangul — kui me just ei tea lisainformatsiooni tegeliku tihedusfunktsiooni kohta — ei õnnestu hinnangu viga vähendada kiiremini kui tuumahinnangul). Tõsi, kui nõustume, et tihedusfunktsiooni hinnang ise ei pruugi olla tihedusfunktsioon (lubame tal omandada ka negatiivseid väärtuseid), on võimalik leida veelgi kiiremini koonduvaid hinnanguid.

Põhimõtteliselt saab leida ka parima tuumafunktsiooni, mille puhul MISE (vähemalt asümptootiliselt) kõige väiksem oleks. Üheks optimaalseks tuumafunktsiooniks on Epanechnikovi tuumafunktsioon. Keskmine võit täpsuses võrreldes teiste tuumafunktsiooni valikutega on aga tühine, tuumafunktsiooni valikut ei peeta üldiselt eriti kriitiliseks probleemiks. Pigem valitakse tuumafunktsioon antud rakenduse jaoks vajalike omaduste järgi (näiteks: kui kasutame normaaljaotust tuumafunktsioonina, siis ei muutu saadud tihedusfunktsiooni hinnang kuskil matemaatilises mõttes nulliks — erinevalt näiteks Epanechnikovi või ühtlasest jaotusest tuleneva tuumafunktsiooni abil saadud hinnangutest. Sõltuvalt ülesandest võib see olla soovitud või ebasoovitav tulemus).

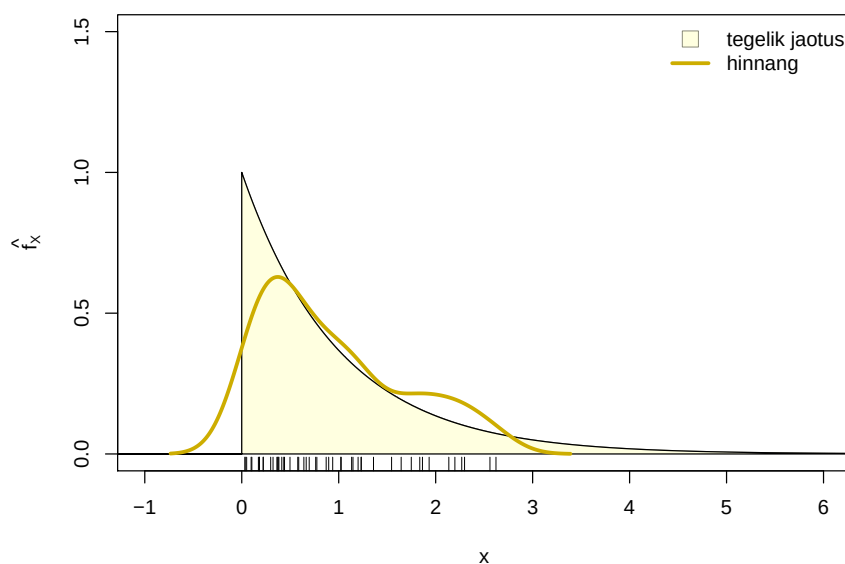
6.2.4 Lisainformatsiooni kasutamisest

Arvuti teab vaid seda informatsiooni, mis tema käsutusse on antud. Vahel arvatakse ekslikult, et piisab statistikapaketile vaid kogutud andmete edastamisest. Enamasti on sellise lähenemise puhul tulemuseks kehvapoolsed tulemused. Kui uurija viitsib mõista, millist tunnust ta uurib, siis avastab ta sageli, et teab tunnuse käitumise kohta märkimisväärselt palju taustainformatsiooni. Näiteks kui uuritavaks tunnuseks on kas kaal või vanus, siis peavad uuritava tunnuse väärtused olema positiivsed, $X > 0$. Kui me arvutiga seda taustainformatsiooni aga ei jaga, hinnatakse andmete pealt sageli tihedusfunktsioon, mis justnagu lubaks ka negatiivsete vaatluste tekkimist, vaata näiteks joonist 6.10.

Uuritava tunnuse kohta olevat taustainformatsiooni tuleks parimate tulemuste saamiseks ka kasutada. Kuidas aga tihedusfunktsiooni hindamisel kasutada teadmist, et uuritava tunnuse väärtused ei saa olla negatiivsed?

Toome siinkohal ära kaks erinevat võimalust lisainformatsiooni $X > 0$ kasutamiseks.

Joonis 6.10: Tihedusfunktsiooni hinnang ilma lisainformatsiooni kasutamata



Tunnuse transformeerimine

Kui $X \in (0 \dots \infty)$, siis $\ln(X) \in (-\infty \dots \infty)$ ehk juhuslik suurus $\ln(X)$ võib paikneda juba reaaltelje mistahes piirkonnas (tema paiknemise kohta meil enam piiranguid pole). Sageli hinnatakse juhusliku suuruse $Y := \ln(X)$ tihedusfunktsioon ja leitakse seejärel juba matemaatiliste teisenduste abil, milline on juhusliku suuruse $X = \exp(Y)$ tihedusfunktsioon.

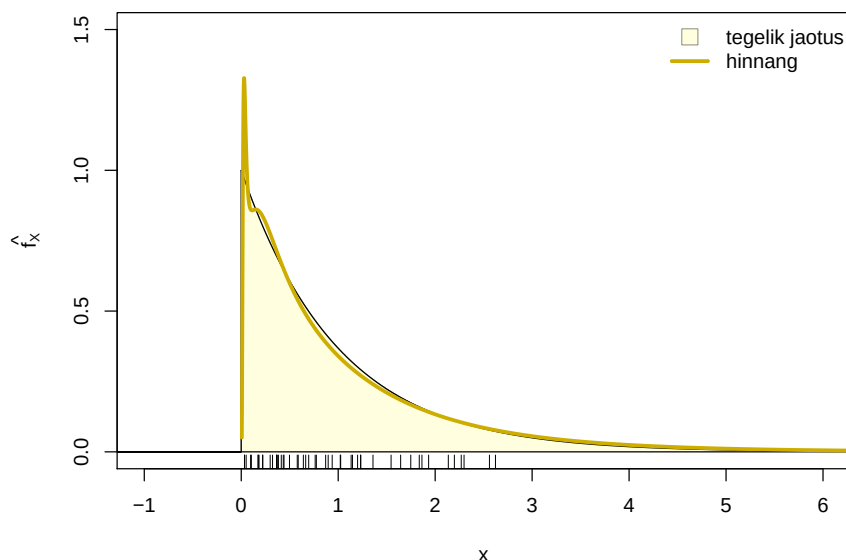
Meeldetuletuseks: Kui soovime leida juhusliku suuruse Y funktsiooni $X = g(Y)$ tihedusfunktsiooni, siis saame vajaliku arvutuse teha järgmise valemi abil (eeldame, et g on funktsiooni määramispiirkonnas I pidev ja monotoonne):

$$f_X(x) = \begin{cases} f_Y(g^{-1}(x)) \cdot |(g^{-1})'(x)|, & x \in I \\ 0 & x \notin I \end{cases}.$$

Kui kasutame transformatsiooni $X = \exp(Y)$, siis $g(x) = \exp(x)$, $g^{-1}(x) = \ln(x)$ ja $(g^{-1})'(x) = 1/x$. Seega $f_X(x) = f_{\ln(X)}(\ln(x)) / x$. Saadud valem võimaldab minna tunnuse logaritmitud väärtustele leitud tihedusfunktsioonilt

üle mittetransformeeritud juhuslike suuruste tihedusele. Antud meetodi rakendamisel saadud tulemust iseloomustab joonis 6.11.

Joonis 6.11: Tihedusfunktsiooni hinnang ($X > 0$) kasutades logaritmilist transformatsiooni



Vaatluste peegeldamine

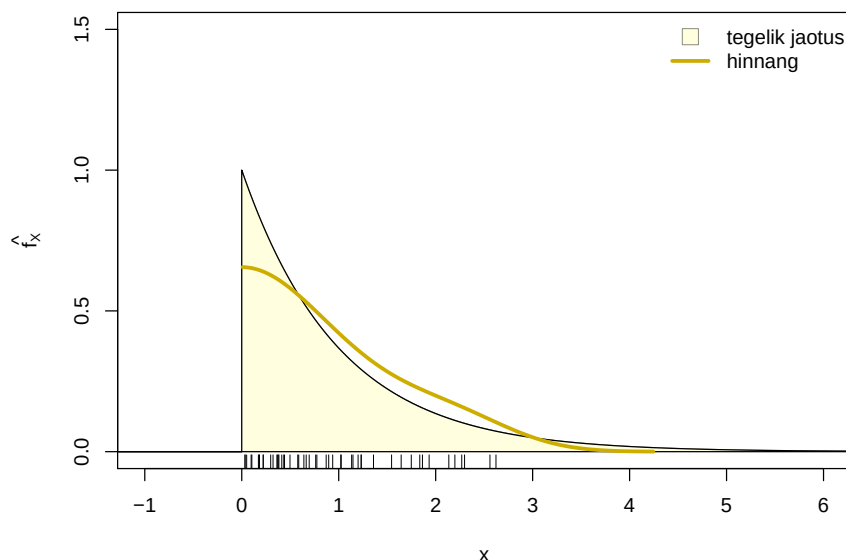
Teine meetod mida kasutatakse kui $X > 0$ kasutab vaatluste peegeldamist nullpunkti suhtes (ja peegeldatud vaatluste lisamises esialgsele valimile). Kui mõõdetud uuritava tunnuse väärtused on näiteks 12, 15, 16, ... siis lisatakse täiendavalt valimile veel vaatlused -12, -15, -16, Saadud uue valimi (mis on esialgsest valimist 2 korda suurem) põhjal leitakse esialgne hinnang tihedusfunktsioonile $\hat{f}_1$. Saadud proto-tihedusfunktsioonist täpselt pool paikneb nüüd negatiivsete väärtuste poolel (juhul, kui kasutame sümmeetrilist tuumafunktsiooni). Tegelikult muidugi ei saa meid huvitav juhuslik suurus X omandada negatiivseid väärtuseid. Seega peab $f_X(x) = 0$, kui $x \leq 0$. Kui võtame ette peegeldatud väärtuseid kasutades saadud proto-tihedusfunktsiooni hinnangu $\hat{f}_1$ ja muudame ta nullist väiksemate x väärtuste korral võrdseks nulliga, siis peame ülejäänud, nullist suuremat poolt tõstma 2 korda (et tu-

lemuseks saadud funktsioon ikkagi tihedusfunktsioon oleks):

$$\hat{f}_X(x) = \begin{cases} 0, & x \leq 0 \\ 2\hat{f}_1, & x > 0 \end{cases}.$$

Vaatluste peegeldamise ehk valimi kahekordistamise meetodil saadud hinnang tihedusfunktsioonile on toodud joonisel 6.12.

Joonis 6.12: Tihedusfunktsiooni hinnang ($X > 0$) kasutades valimi kahekordistamise ehk vaatluste peegeldamise meetodit



Kumb kahest väljapakutud meetodist paremini töötab, sõltub konkreetsest andmestikust. Logaritmilist transformatsiooni kasutava meetodi puhul on tihedusfunktsiooni hinnangu parempoolne piirväärtus nullpunktis alati 0 ($\lim_{x \rightarrow 0+} \hat{f}_X(x) = 0$) ehk saadud tihedusfunktsiooni hinnang on pidev punkti $x = 0$ ümbruses. Seevastu väärtuste peegeldamise meetodi puhul sellist nõuet pole, tihedusfunktsiooni hinnangul esineb sageli katkevuspunkt kohas $x = 0$.

6.2.5 Rakendusnäide

Üheks sageli statistikule ettetulevaks ülesandeks on klassifitseerimisülesanne: vaadeldud tunnuse või tunnuste põhjal tuleb uuritavad objektid jagada kahte või enamasse teadaolevasse klassi (pank soovib kliente jagada laenu tagasimaksvateks ja makseraskustesse sattuvateks klientideks; geneetik soovib mõne välise tunnuse järgi aru saada, kas uuritav on AA, AB või BB genotüübiga; botaanik soovib heinakörre laiuse või mõne muu tunnuse põhjal määrata mis liiki taimega on tegemist jne). Tavaliselt on sellise ülesande lahendamiseks kasutada valim, kus eksperdi poolt või kalleid analüüse kasutades on määratud objektide klassikuuluvus. Statistiku ülesandeks on valimi eeskujul lahterdada ka uued objektid etteantud klassidesse.

Jagagem uuritavaid objekte k -sse erinevasse klassi. Oletame, et on teada iga klassi esinemistõenäosus uuritavas populatsioonis $(p_1, p_2, \dots, p_k)$ ja teame ka klassifitseerimiseks kasutatava tunnuse (või tunnuste) jaotust iga klassi jaoks eraldi — teame, milline on tunnuse X jaotus liiki A kuuluvate taimede korral, milline on tunnuse X jaotus liiki B kuuluvate taimede jaoks jne ehk eeldame, et antud on tihedused $f_1 := f_{X|klass=1}, f_2 := f_{X|klass=2}, \dots, f_k$. Kui peaksime nüüd klassifitseerima uue objekti, mille puhul $X = x$, siis võiksime esmalt leida tinglikud tõenäosused $P(\text{objekt kuulub klassi } 1|X = x), \dots, P(\text{objekt kuulub klassi } k|X = x)$ kasutades tingliku tõenäosuse valemit:

$$P(\text{objekt kuulub klassi } i|X = x) = \frac{f_i(x) \cdot p_i}{f_X(x)},$$

kus f_X on tunnuse X jaotus uuritavas populatsioonis (üle kõigi klasside):

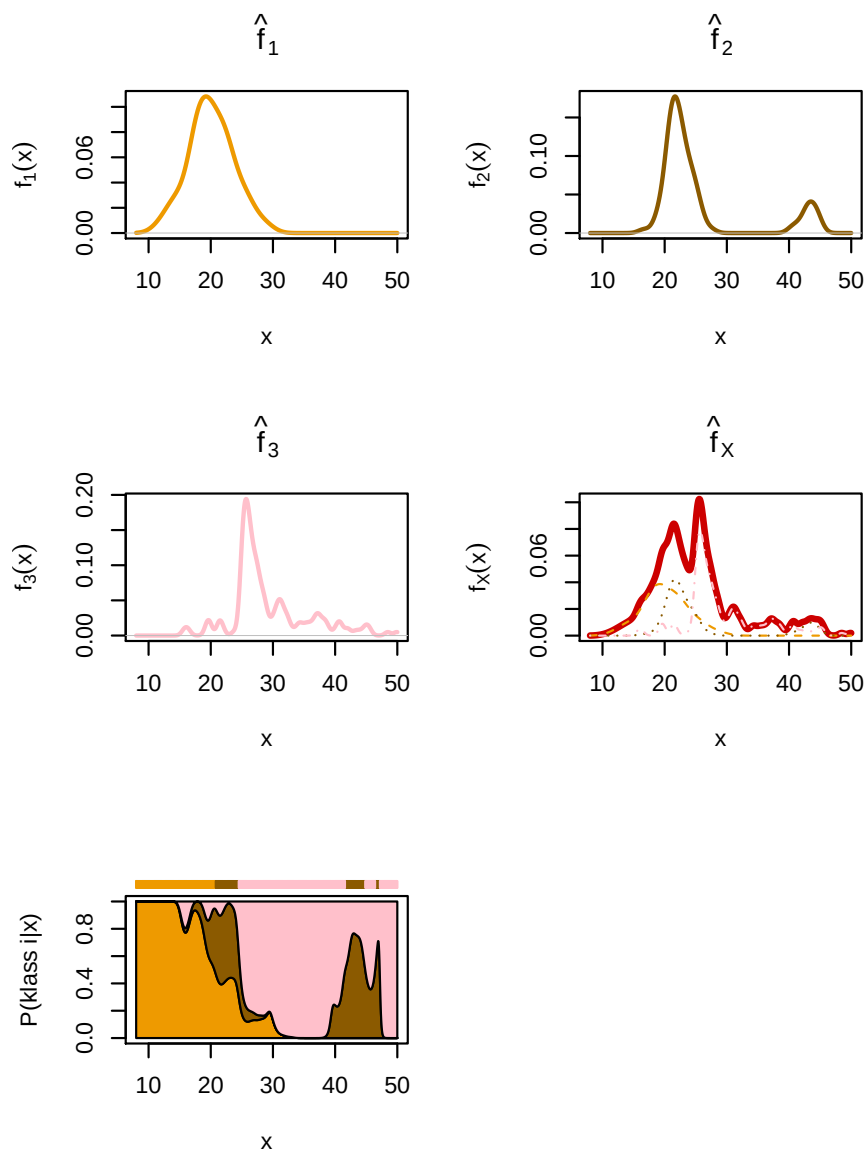
$$f_X(x) = f_1 p_1 + f_2 p_2 + \dots + f_k p_k.$$

Kasutades leitud tinglikke tõenäosuseid on aga võimalik uut uuritavat klassifitseerida — näiteks määrates ta kõige tõenäolisemasse klassi.

Loomulikult pole reaalse ülesande korral võimalik kasutada ülesande lahendamiseks tõenäosuseid $p_1, \dots, p_k$ ega ka tihedusfunktsioone $f_1, \dots, f_k$, küll aga on võimalik mõlemaid suuruseid hinnata uurija käsutuses oleva valimi põhjal. Saadud hinnanguid kasutades on võimalik leida ka hinnanguid tinglikele tõenäosustele.

Vaata ka joonist 6.13.

Joonis 6.13: Klassifitseerimisülesande lahendamine tihedusfunktsiooni hinnanguid kasutades



tihedusfunktsiooni hinnang tüüpiliste valimi suuruste juures väga ebatäpseks (talutava täpsuse saavutamiseks võime vajada väga suurt valimit). Selliseid sobivaid lisaeelduseid on aga sageli mugavam sisse tuua mõne teise tihedusfunktsiooni hindamiseks mõeldud meetodi korral.

Sisukord

1	Sissejuhatus	3
1.1	Mitteparameetrilise statistika mõistest	3
1.2	Tihedusfunktsiooni hindamine	4
1.3	Märgitest	5
1.4	Usaldusintervall mediaanile	7
1.5	Mediaani mitteparameetriline hindamine	8
2	Wilcoxon astakmärgitest	9
2.1	Motivatsioon ja teststatistik	9
2.2	Statistiku W jaotus nullhüpoteesi kehtides	10
2.2.1	Monte-Carlo meetod olulisustõenäosuse arvutamiseks	13
2.3	Asümptootiline jaotus	13
2.3.1	Pidevuse parandus	16
2.4	Võrdsed astakud	17
2.5	Erinevuse kirjeldamine	19
3	Wilcoxon astaksummatest	21
3.1	Teststatistik	21
3.2	Keskvärtus ja dispersioon H_0 kehtides	24
3.3	Asümptootiline jaotus	27
3.4	Populatsiooni mudelist	29
3.5	Testi võimsus	30
3.5.1	Keskvärtus ja dispersioon H_1 kehtides	30
3.5.2	Asümptootiline võimsus	32
3.5.3	Wilcoxon ja t-testi võrdlus	37
3.6	Wilcoxon astaksummatesti tulemuse tõlgendamisest	39
3.7	Võrdsed vaatlused	41
3.8	Siegel-Tukey ja Ansari-Bradley testid	41
	Ülesanded	44

4	Friedmani test	45
4.1	Friedmani teststatistiku jaotus H_0 kehtides	46
4.2	Asümptootiline jaotus nullhüpoteesi kehtides	47
4.3	Friedmani test korduvate väärtuste olemasolu korral	49
4.4	Sobitatud juht-kontrolluuring — McNemar'i test	50
4.5	Post-hoc testid	52
4.6	Ajalooline vahepala	53
5	Kruskal-Wallise test	55
5.1	Kruskal-Wallise statistiku jaotus H_0 kehtides	56
5.2	Statistiku asümptootiline jaotus H_0 kehtides	57
5.3	Veel mõningaid tulemusi	60
6	Tihedusfunktsiooni ja jaotusfunktsiooni hindamine	61
6.1	Jaotusfunktsiooni hindamine	61
6.2	Tihedusfunktsiooni hindamine	63
6.2.1	Tihedusfunktsiooni hindamine empiirilise jaotusfunktsiooni abil	65
6.2.2	Tuumameetod tihedusfunktsiooni hindamiseks	67
6.2.3	Hinnangu teoreetilised omadused	73
6.2.4	Lisainformatsiooni kasutamisest	74
6.2.5	Rakendusnäide	78
6.3	Mitmemõõtmeline tihedus	80