TARTU ÜLIKOOL

Loodus- ja tehnoloogiateaduskond

Tehnoloogiainstituut

Hannes Luidalepp

Teaduskonna kümnevõistluse tulemuste analüüs

„Biomeetria bioloogidele“ iseseisev andmeanalüüsi töö

Tartu, 2010

SISUKORD

SISSEJUHATUS

Tartu Ülikooli Bioloogia-geograafia teaduskonnas ja selle järglases Loodus- ja tehnoloogiateaduskonnas on juba 10 korda toimunud kümnevõistlus. Sellest meeleolukast võistlusest on osa võtnud suur hulk erinevaid sportlasi, kellest mõni on osalenud korduvalt, mõni ainult ühe korra. Osalejate vorm on samuti väga erinev. Mõned võistlejad teevad sporti ainult korda aastas, siis kui toimub järjekordne teaduskonna mitme võistlus, paljud spordivad regulaarselt ning mõningad on pea, et profisportlased treenides igapäev.

Kuna sportlaste tase on väga erinev ning nende spordiharjumised samuti, on minul, kui ühel korraldaja,l tekkinud küsimus, kes meil ikkagi võistlemas käivad? Mind huvitab, kas osalejad on esinduslik valim populatsioonist (Eesti meessoost 20-30 aastased elanikud) või saab eristuvad võistlejate hulgas mingid kindlad populatsioonid (nt tõsised harrastajad ja asportlikud naljamehed).Samuti oleks huvitav teada, kas mõne üksikala järgi oleks võimalik enam-vähem ennustada võistleja kogutulemust. Selliseid statistika abil vastatavaid küsimusi ongi plaanis käesolevas andmeanalüüsitöös käsitleda.

1. ANDMESTIKU KIRJELDUS

Käes oleva andmeanalüüsitöö andmestik koosneb kõigi ajajooksul toimunud Lote (endise BioGeo) kümnevõistluste tulemustest ning punktidest. Kokku on esindatud 182 võistlemist, kus juures iga võistlemise juures on eraldi välja toodud võistleja nimi, üksikalade tulemus ja punktid, punktide kogusumma ning võistlemise aasta. Vahele jäetud alade eest ei ole punkte üldse antud (ka 0 punkti ei ole antud). Esindatud on 71 erinevat võistlejat, kes minimaalselt on osalenud 1 võistlusel ning maksimaalselt 9 võistlusel. Võistlejateks on teaduskonna meessoost üliõpilased, endised üliõpilased, mõni üksik töötaja ning lisaks ka mõned üksikud nn väliskülalised. Kuna sportlasi on erineva vanusega, erinevatelt kursustelt ning ka erinevatelt erialadelt võiks valim olla eesti 18-30 aastaste meeste populatsiooni suhtes üsnagi esinduslik.

Üksikaladel on välja toodu nii tulemus (meetrid või sekundi) kui ka punktid. Punktid on tulemustega seotus läbi astme funktsiooni, mida tuleb arvestada, kui omavahel võrrelda punkte ja tulemusi. Lisaks on mõnel alal minimaalne tulemus (jooksud) punktide saamiseks üsna kõrge, nii et paljud kehvemas vormis võistlejad ei saagi punkte. Tagajärjeks on selliste alade punkti ja tulemuse jaotuse suur erinevus.

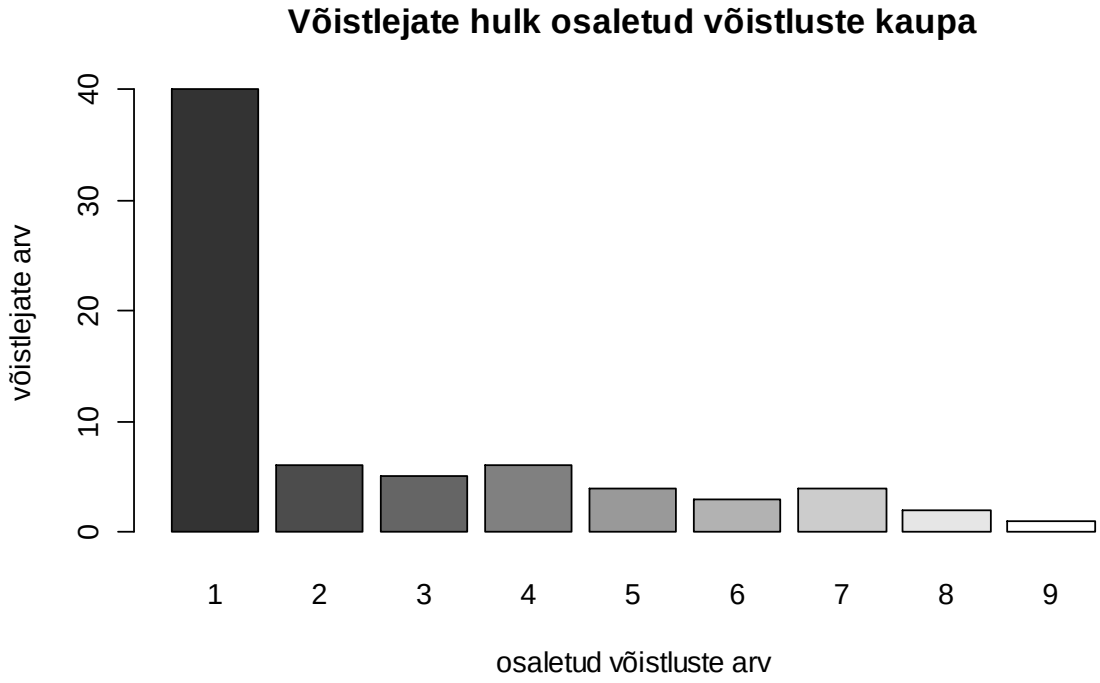
2. ANDMESTIKU ANALÜÜS

Võistlejad

Esimese asjana tahaks teada, kes on võistlejad? Selleks vaatasime, mitmest võistlusest olid erinevad võistlejad osa võtnud (Joonis 1). Üllataval kombel selgus, et enamus võistlejad on osalenud ainult ühe korra ning võistlejate hulk, kes on osalenud enamal kordadel on üsnagi ühtlane (negatiivse astendajaga astme funktsioon?). Samas eeldasin, et võistlustest osavõtmiste arv on ühtlasem. Nende andmete valguses tundub, et on kahte sorti inimesi: ühed kes, proovivad korra ja rohkem ei tule, ning teised kes korra proovivad ning jäävadki niikauaks käime kuni neil see on võimalik (nt. lõpetavad ülikooli). Ühe-korra võistlejate suure hulga põhjustab ilmselt kooli ja võistluse korralduse võimalik vastu olu. Nimelt toimub võistlus enamasti juuni alguses või mai keskel, samas ajal on inimestel vaja eksameid teha (esmakursuslased ei julge ühtegi vaba päeva võtta) või siis lõputööd kirjutada (bakalaureuse 3. kursus ja magistri 1. kursus). Nii võibki vähem innustunud üliõpilasel võistluseks üldse 2 kevadet sobida, millest ehk ühe ajal on muudki vaja teha.

Tulemused

Sportlikud tulemused sõltuvad kahest olulisest faktorist, need on sooritusvõime ning tehnika, kusjuures nende osakaalud on erinevatel aladel erinevad. Harrastaja/mitte-sportiva inimese tasemel võib sooritusvõime või tehnika puudujääki kompenseerida teataval määral kehalised eeldused (nt kehakaal, pikkus). Seetõttu on erinevate alade lõpptulemus segu erinevatest näitajatest ning vastavalt spordiala spetsiifikale võib tulemuste jaotus olla väga erinev. Seetõttu vaatasingi järgmisena, milline on üksik alade tulemuste jaotus- kas tulemused on



Joonis 1. Võistlejate hulk osaletud võistluste kaupa.

ühtlased ning jaotuvad vastavalt mõnele normaaljaotusele või esineb neis mõni spetsiifiline muster (Joonis2, 3 ja 4).

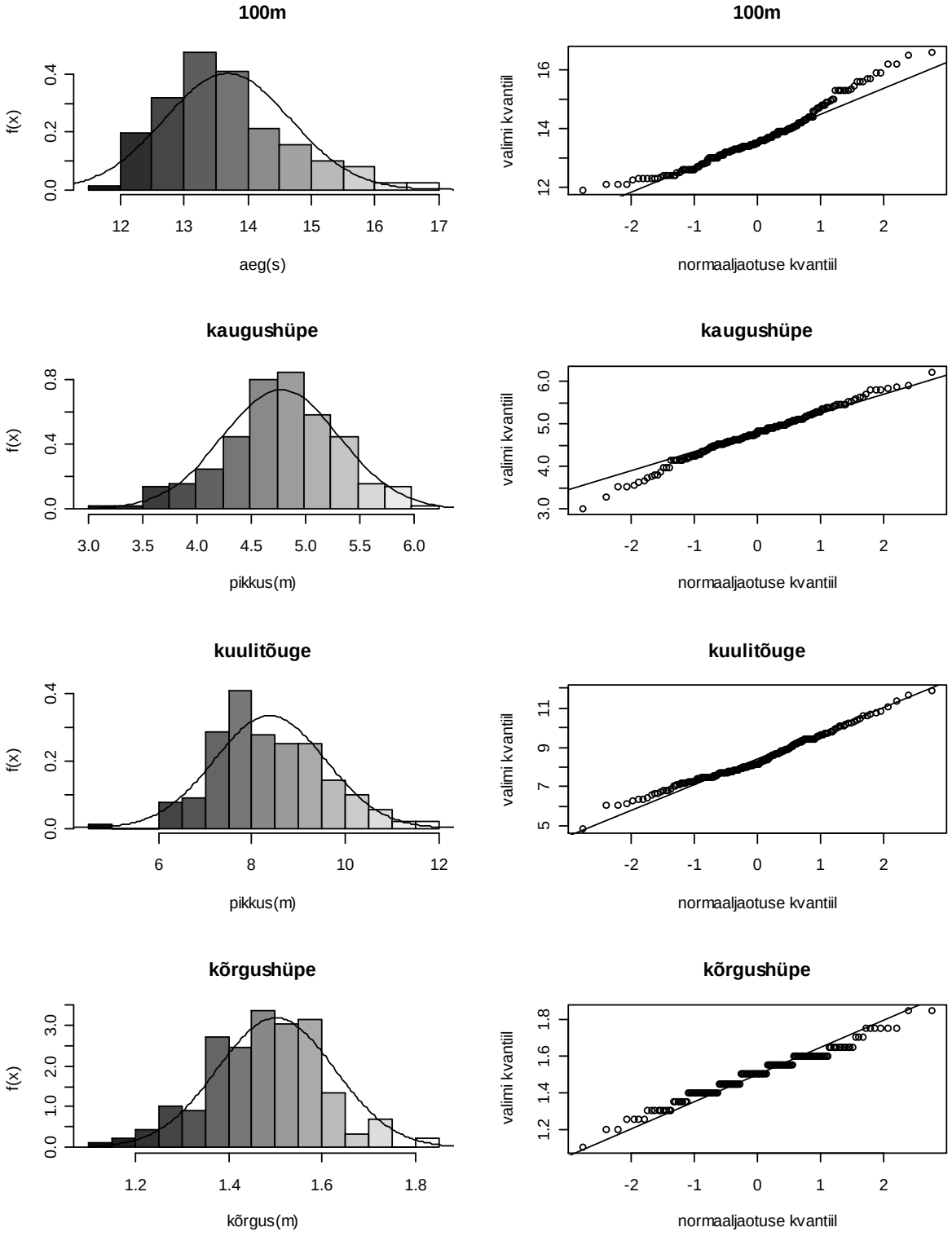
Kõige huvitavama jaotusega on 110mtj, mis tundub koosnevat kahest, võimalik et normaaljaotusega, populatsioonist. Kuna tõkkejooks on üsna tehniline ning nõuab parasjagu füüsist, siis tundub, et võistlejad, kelle on nõrk füüsiline ja tehniline ettevalmistus võistlejad moodustavad eraldi populatsiooni. See tähendab, et füüsise või tehnika valdamine võib anda mitte-valdajate ees olulise eelise ning kehalised eeldused olulist rolli ei mängi.

Ülejäänud jooksud (100m, 400m, 1500m) kippusid huvitaval kombel olema ühes suunas välja venitatud jaotusega, mis tuleneb ilmselt sellest, sportlikumate võistlejate omavahelised erinevused on väiksemad kui vähem sportlike omad. Ühtlase jaotuse tagab nende alade juures ilmselt suur sooritusvõime roll tulemuses (tehnika suhteliselt väike) ning sooritusvõime ühtlane jaotus võistlejate vahel.

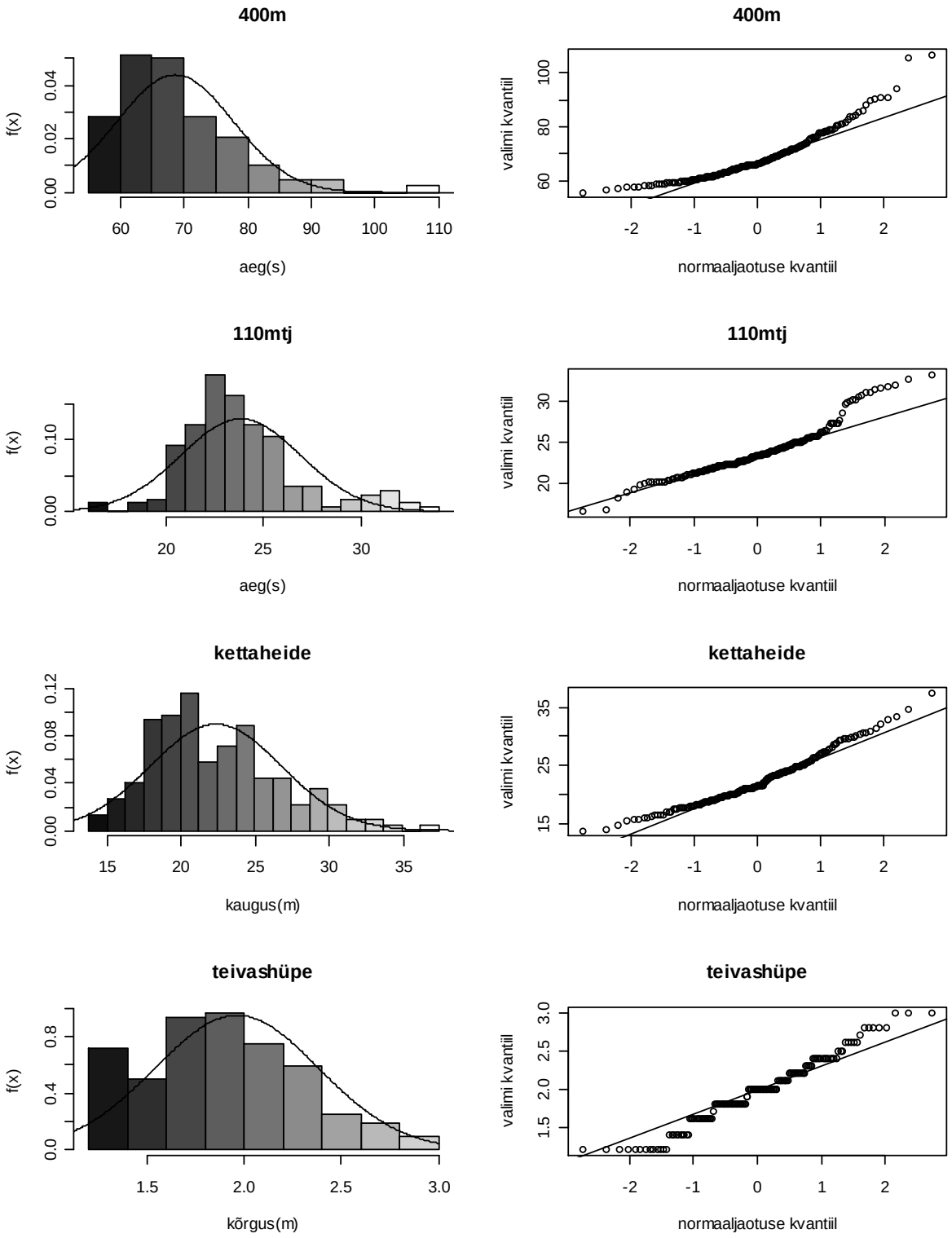
Kaugushüppe, kõrgushüppe ning kuulitõuke tulemuse tunduvad olevat normaaljaotusega. Kuigi ka need alad on tehnilised, aitab harrastaja tasemel tehnilist või füüsilist puudujääki korvata suurem kehakaal (kuulitõuge) või pikkus (kaugus- ja kõrgushüpe), mis võib tagada tulemustes normaaljaotuse. See tähendab, et ainult tehnika valdamine või parem füüsiline ettevalmistus ilmselt ei taga nii selget eelist võistlejate ees, kes on nõrgad nii tehnikas kui ka füüsises.

See-eest ei ole normaaljaotusega teivashüpe, kettaheide ning odavise, mis võivad samuti koosneda mitmest erinevast normaaljaotusega populatsioonist. Viimased kolm ala on see-eest sellised, kus tehnilist või soorituslikku vajaka jäämist kehalise omadused eriti ei asenda, mistõttu tulemused normaaljaotusele ehk ei vastagi. See tähendab, et parema tehnilise ettevalmistusega võistleja võib saada teiste ees parema eelise. Seejuures on huvitav, et kui sooritusvõime aladel (jooksud) on tulemused ühtlased, kuid tehnika aladel, kus kehalised

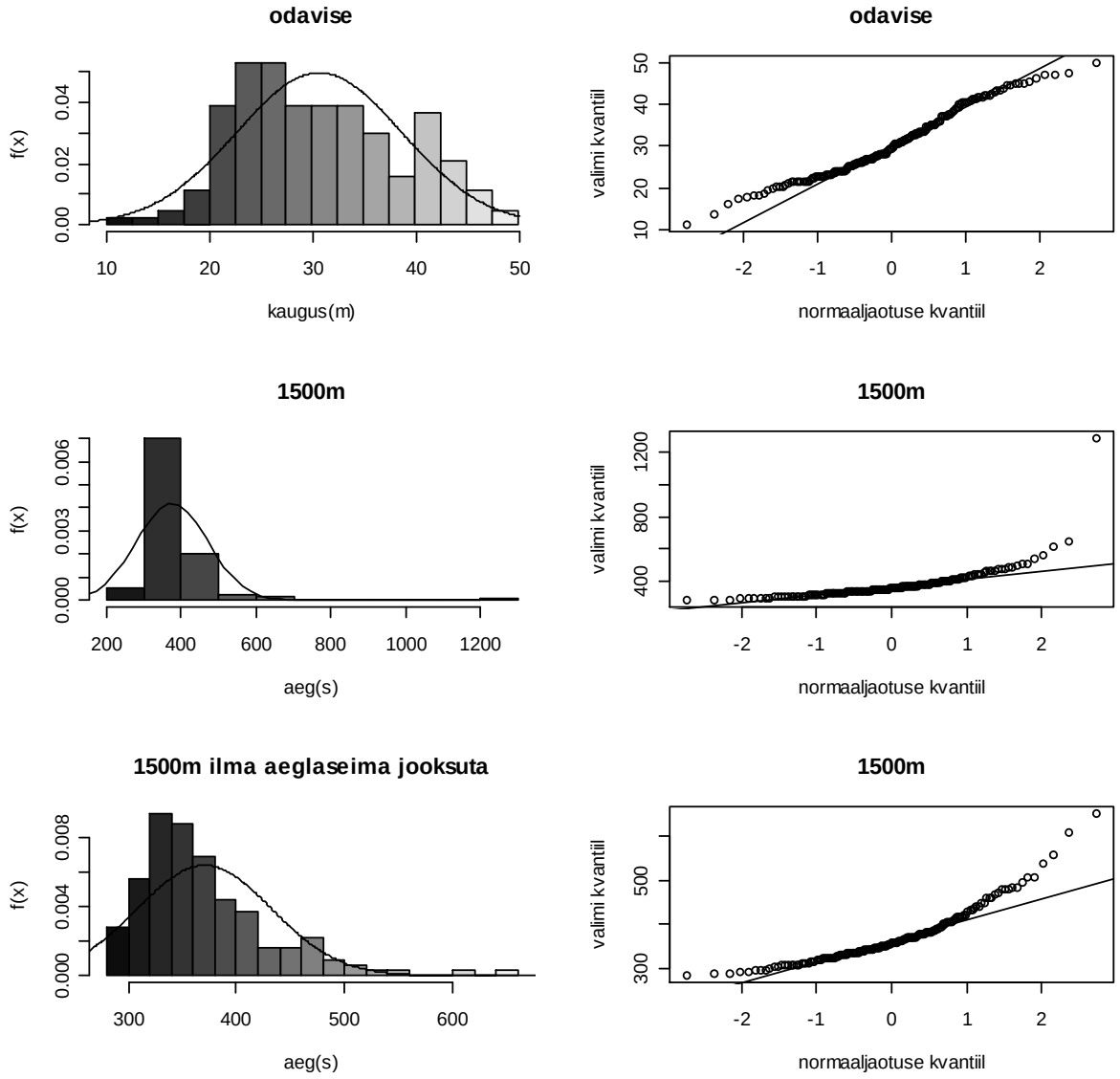
eeldused eriti kasuks ei tule (teivashüpe, kettaheide ning odavise) võivad tulemused olla segu mitmest populatsioonist. Järeldus oleks, et tehniline areng võib anda hüppelisi parandusi tulemustes.



Joonis 2. Erinevate alade tulemuste jaotus koos kõige paremini sobiva normaaljaotusega ning tõenäosuspaberiga I.



Joonis 3. Erinevate alade tulemuste jaotus koos kõige paremini sobiva normaaljaotusega ning tõenäosuspaberiga II.



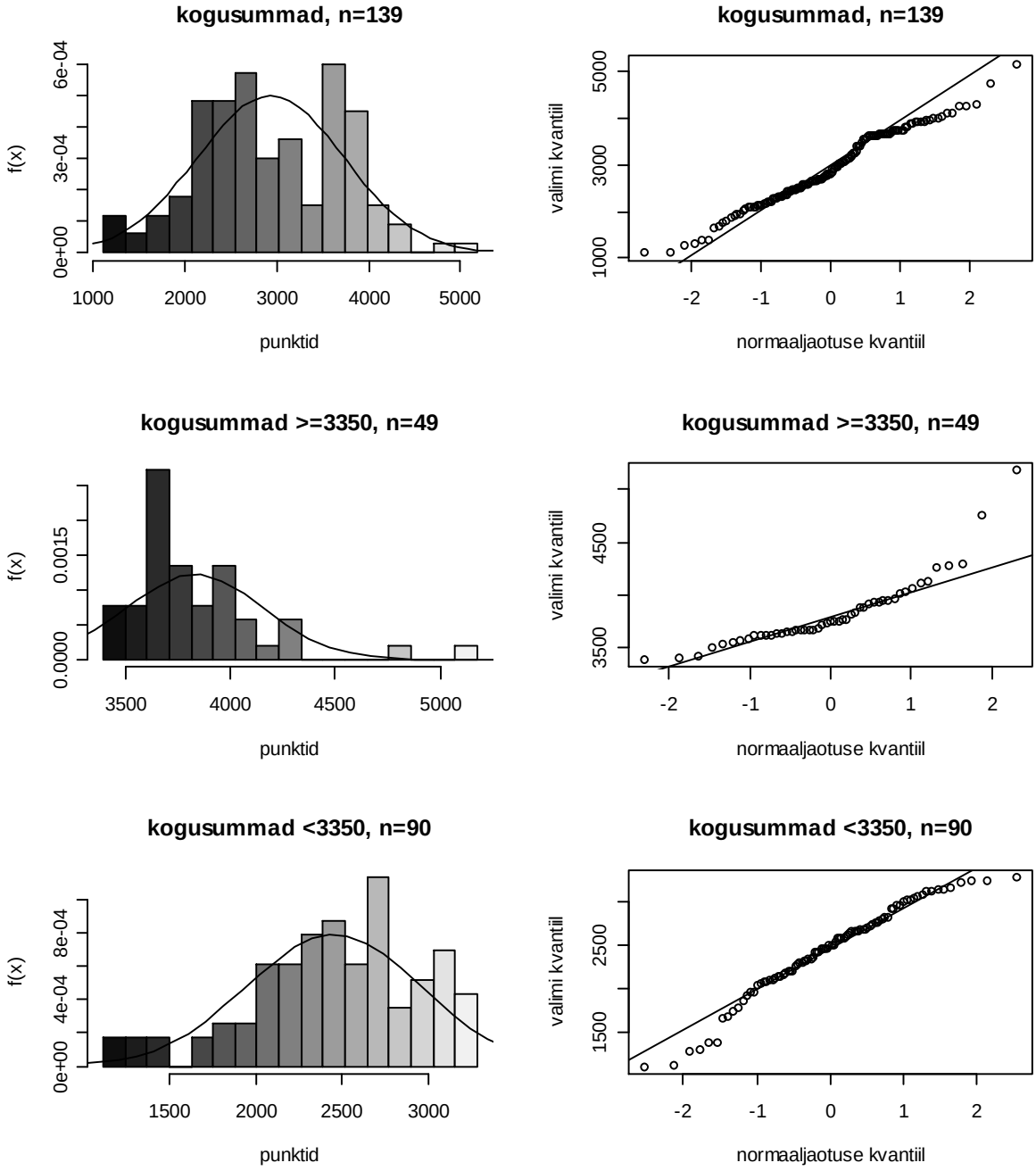
Joonis 4. Erinevate alade tulemuste jaotus koos kõige paremini sobiva normaaljaotuse ning tõenäosuspaberiga III.

Punktid

Lisaks huvitab mind, kuidas on jaotunud punktid, kuid neid ei hakka ruumi ja ajapuudusel üks haaval välja tooma. Lisaks muudab nende interpreteerimise raskeks, punktide ja tulemuste seos läbi astmefunktsiooni ning mõningatel aladel on minimaalne punkte toov tulemus suhteliselt kõrge, mis tõttu paljud võistlejad ei saa üldse punkte.

Seetõttu otsustasin punktidest vaadata ainult kogusumma jaotumist (Joonis 5). Kuna 181 võistlemisest on osad katkestatud ning mõni ala vahele jäetud otsustasin vaadata kogusumma jaotumist ainult nendel võistlemistel, mis on täielikult sooritatud (139 võistlemist). Selgub, et see jaotus ei vasta normaaljaotustele, pigem tundub, et see jaotus koosneb kahest eraldi populatsioonist, mis võiksid esindada erineva tasemega võistlejaid. Järgmisena jagasin andmestiku 3350 punkti pealt kaheks ja võrdlesin kahe andmestiku jaotusi (Joonis 5). Jagatud

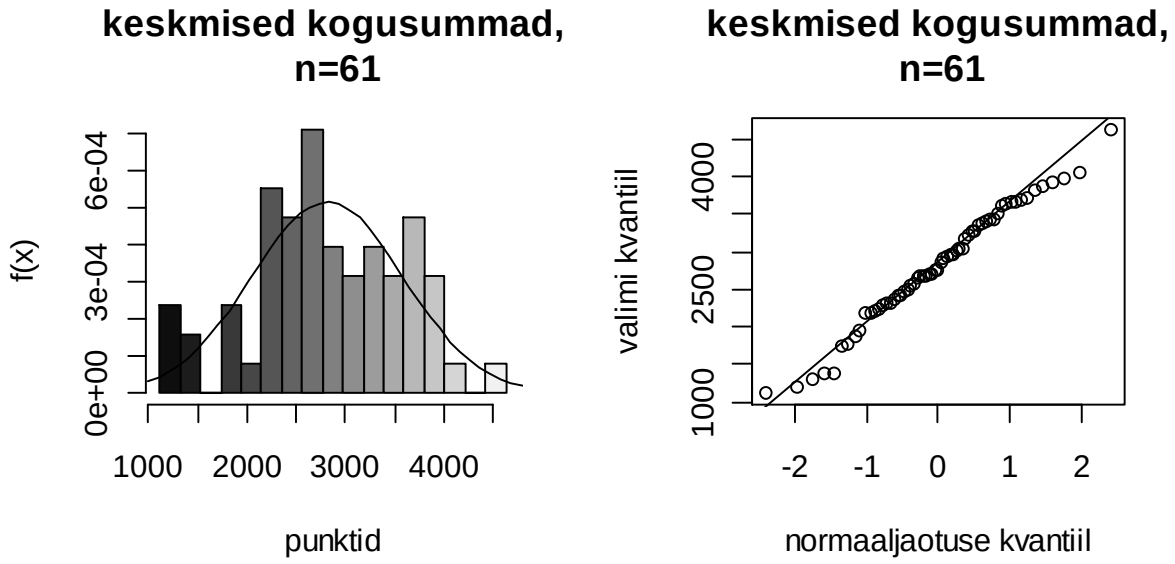
andmestiku osad ei andnud eraldi samuti päris head vastavust normaaljaotusele, kuid jaotus oli normaaljaotusele juba lähemal. See võib tähendada, et meil osaleb võistlustel tõesti kahe eristuva tasemega võistlejaid (nt. ühed on kergejõustiku trennis käinud, teised ei ole).



Joonis 5. Täielikult sooritatud võistlemiste punktide kogusumma jaotus koos kõige paremini sobiva normaaljaotuse ning tõenäosuspaberiga. Lisaks on punktide kogusumma jagatud kaheks andmestikuks, koos kõige paremini sobiva normaaljaotuse ning tõenäosuspaberiga.

Kuna hulk võistlejad on osalenud mitmeid ning mõned ainult ükikutel kordade, siis ka see võib jaotust muuta (tegemist oli kõikide tulemuste jaotusega). Et seda välistada ka võtsin täielikult sooritatud võistluste keskmised võistlejate kaupa (61 võistlejat) ja vaatasin seda

jaotust (Joonis6). See jaotus vastab juba paremini normaaljaotusele (arvestades üsna väikest valimi suurust) või vähemalt on tegemist üsna ühtlase jaotusega. Seega ei julge väita, et meie võistlusest võtaks osa ainult mingid kindlad eristuvate sportlike võimetega sportlased. Pigem on osalejad ikkagi küllaltki ühtlaselt erineva sportliku tasemega inimesed.



Joonis 6. Võistlejate keskmised täielikult sooritatud võistlemiste kogusummad koos kõige paremini sobiva normaaljaotuse ning tõenäosus paberiga.

3. SEOSED ANDMESTIKUS

Tippkümnevõistlejad ütlevad, et 100m jooksu tulemuse järgi on kõige parem ennustada lõpptulemust. Võimalik, et sellel on ka statistiline tagapõhi, kuna 100m jooks näitab plahvatuslikkust ning kiirust, mis tippkümnevõistlejate puhul on kõige olulisem.

Samas huvitab mind, millise ala või alade järgi on kõige parem ennustada nö harrastaja tasemel kümnevõistluse tulemust. Selleks vaatasin erinevate alade tulemuste ja punktide korreleerumist kogusummaga (Tabel 1). Selgub, et kõige praemini korreleerusid kaugushüpe ning 400m jooks (eriti just nende punktid). Kuid mõlema ala punktide järgi regressiooni mudel ei olnud väga suure ennustusliku väärtusega (R^2- väärtused olid vastavalt 0,7308 ja 0,6644). Seetõttu otsustasin luua mudeli, mis kasutaks kaugushüppe ja 400m jooksu punkte. Selle mudeli determinatsiooni kordaja on 0,8539, mis näitab, et uus mudel on palju suurema ennustusliku väärtusega. Mudel ise on toodud tabelis 2.

Mudeli kontrollimiseks peab mudel vastama mõningatele eeldustele ja neid hakkangi järgnevalt kontrollima.

Tabel 1. Lõpptulemus korreleerumine üksikalade punktide ja tulemustega

X100m	**-0.7001412**
X100m_p	**0.7247731**
kaugus	**0.8542905**
kaugus_p	**0.8548908**
kuul	**0.5211950**
kuul_p	**0.5205004**
kõrgus	**0.6200573**
kõrgus_p	**0.6132801**
X400m	**-0.7890898**
X400m_p	**0.8150824**
X110mtj	**-0.7209445**
X110mtj_p	**0.7429418**
ketas	**0.6255331**
ketas_p	**0.6241615**
teivas	**0.6627815**
teivas_p	**0.6630652**
oda	**0.6662156**
oda_p	**0.6649842**
X1500m	**-0.5572850**
X1500m_p	**0.7401494**
kokku	**1.0000000**

Tabel 2. Täiustatud mudel kümnevõistluse lõpptulemuse hindamiseks.

Kogu summa = 753,139 + 2,396*x + 4,582*z + ε

x = 400m jooksu punktid z = kaugushüppe punktid ε = prognoosi viga

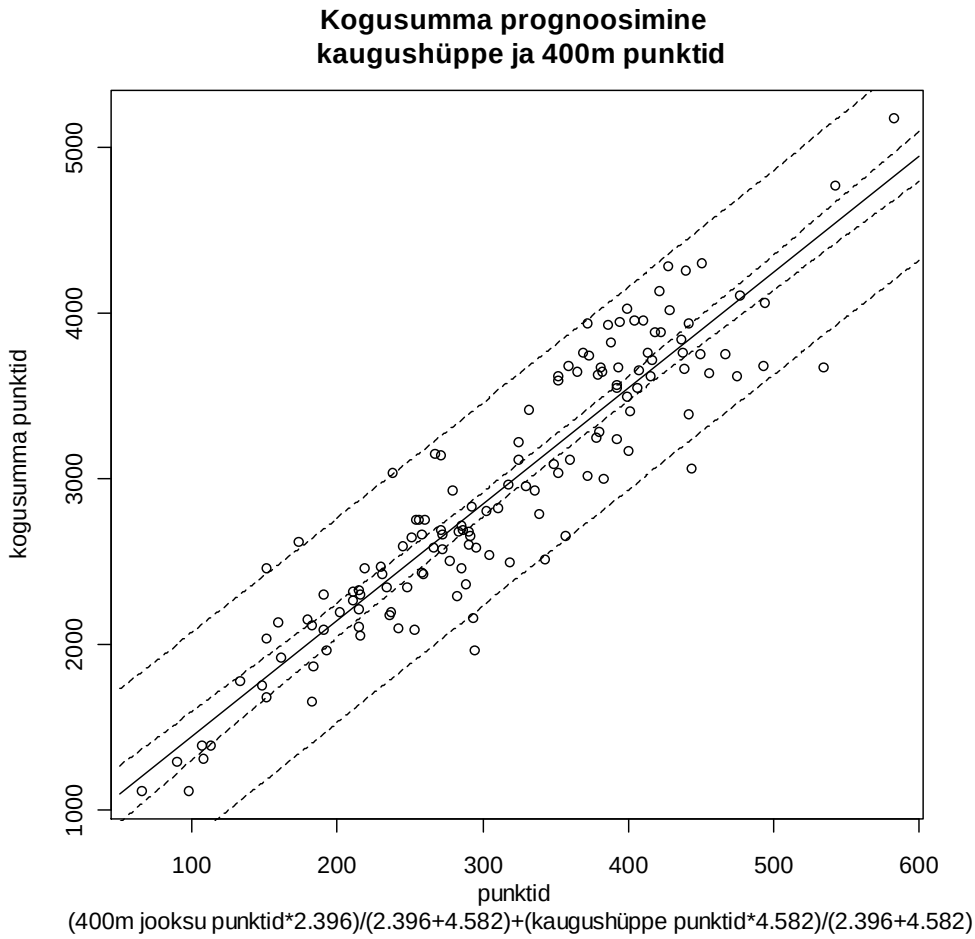
Esiteks tuleb kontrollida, kas sirge ikka sobib seost iseloomustama. Joonisel 8 on esitatud mudel koos käesolevas tööskasutatud andmestiku andmetega ja selle joonise järgi tundub, et lineaarne seos sobib antud mudelisse.

Teiseks peavad mudeli jäägid olema ligikaudu normaaljaotusega ning lisaks peab jääkide hajuvus olema ligikaudu konstantsed. Ka need eeldused on enam-vähem täidetud.

Lisaks hindaks selle mudeli puhul ka rakenduslikku väärtust. Esiteks on positiivne, et mudelisse on kaasatud alad, mida lihtne läbi viia ning mis ei vaja väga erilist varustust (nt teivas). See on oluline, et seda mudelit saaks kasutada nö ennetuslikult, et saada juba enne võistlust hinnangut oma tulemustele. Samas võib tekkida küsimus, kas mitmevõistlusest pärinevate tulemuste põhjal loodud mudelit saab ikka kasutada nö ennetuslikult. Arvan, et saab, kuna mõlemad alad on esimese päeva alad ning nende alade tulemused ei tohiks veel

kannatada koguneva väsimuse tõttu. Kaugushüpe, mis omab mudelis suuremat kaalu, on juba teine ala ning seal ei tohiks väsimusest veel juttugi olla.

Lisaks näitab kaugushüpe kiirust ning plahvatuslikkust ning 400m jooks kiiruslikku vastupidavust, mis kõik on olulised komponendid mitmevõistleja füüsilisest võimekusest.



Joonis 8. Kogusumma prognoosimine kaugushüppe ja 400m jooksu tulemuste järgi vastavalt mudelile (Kogu summa = 753,139 + 2,396*x + 4,582*z + ε). Lisatud on ka prognoosiintervallid ja usaldusintervallid.

KOKKUVÕTE

Kokkuvõtvalt ütleks, et Lote (endine BioGeo) mitmevõistluse osavõtjaskonna näol on tegemist üsna ühtlaselt erinevate sportlike võimetega meesisikutega.

Lisaks on leitud lineaarne mudel, mis erineva tasemega harrastussportlaste mitmevõistluse kogusummat suudab enamvähem rahuldava täpsusega hinnata, võttes aluseks kaugushüppe ning 400m jooksu tulemused punktidena.

LISA 1 – PROGRAMM

```
#Andmete lugemine R-i
a10v=read.csv2("V:/Hannes/10v/10v.csv", header=T)

#Andmestiku kasutamiseks määramine
attach(a10v)

#Võistluste hulk (ridade arv andmestikus)
length(nimi)

#Erineva nimega sportlaste hulk
length(table(nimi))

#Minimaalselt ja maksimaalselt osaletud võistlusi.
range(table(nimi))

#Võistlejate hulk osaletud võistluste kaupa
barplot(table(table(nimi)),
main="Võistlejate hulk osaletud võistluste kaupa",
ylab="võistlejate arv", xlab="osaletud võistluste arv",
col=gray(2:10/10))

#Joonised kaheksa kaupa
par(mfrow=c(4,2))

#100m tulemuste jaotus
hist(X100m,freq=F
,main="100m",
ylab="f(x)", xlab="aeg(s)",
col=gray(1:11/11))
x=seq(0,20, length=500)
y=dnorm(x,mean=mean(X100m, na.rm=T), sd=sd(X100m, na.rm=T))
lines(x,y,lwd=1.5)
qqnorm(X100m,main="100m",        ylab="valimi        kvantiil",
xlab="normaaljaotuse kvantiil")
qqline(X100m)

#kaugushüppe tulemuste jaotus.
hist(kaugus,
breaks=seq(min(kaugus,na.rm=T),max(kaugus,na.rm=T),length=14),
freq=F,main="kaugushüpe",
ylab="f(x)", xlab="pikkus(m)",
col=gray(1:13/13) )
x=seq(0,20, length=500)
y=dnorm(x,mean=mean(kaugus, na.rm=T), sd=sd(kaugus, na.rm=T))
lines(x,y,lwd=1.5)
qqnorm(kaugus,main="kaugushüpe",    ylab="valimi    kvantiil",
xlab="normaaljaotuse kvantiil" )
qqline(kaugus)
```

```
#kuulitõuke tulemuste jaotus.
hist(kuul,freq=F,breaks=10, main="kuulitõuge",
ylab="f(x)", xlab="pikkus(m)",
col=gray(1:15/15) )
x=seq(0,20, length=500)
y=dnorm(x,mean=mean(kuul, na.rm=T), sd=sd(kuul, na.rm=T))
lines(x,y,lwd=1.5)
qqnorm(kuul,main="kuulitõuge"    ,    ylab="valimi    kvantiil",
xlab="normaaljaotuse kvantiil")
qqline(kuul)

#kõrgushüppe tulemuste jaotus ilma 0 tulemuseta.
hist(kõrgus, breaks=20, freq=F,main="kõrgushüpe",
ylab="f(x)", xlab="kõrgus(m)",
col=gray(1:15/15) )
x=seq(0,2, length=500)
y=dnorm(x,mean=mean(kõrgus, na.rm=T), sd=sd(kõrgus, na.rm=T))
lines(x,y,lwd=1.5)
qqnorm(kõrgus,main="kõrgushüpe",    ylab="valimi    kvantiil",
xlab="normaaljaotuse kvantiil" )
qqline(kõrgus)

#400m tulemuste jaotus.
hist(X400m,freq=F,main="400m",
ylab="f(x)", xlab="aeg(s)",
col=gray(1:11/11) )
x=seq(40,120, length=500)
y=dnorm(x,mean=mean(X400m, na.rm=T), sd=sd(X400m, na.rm=T))
lines(x,y,lwd=1.5)
qqnorm(X400m,main="400m"     ,     ylab="valimi     kvantiil",
xlab="normaaljaotuse kvantiil")
qqline(X400m)

#110mtj tulemuste jaotus.
hist(X110mtj,breaks=19,freq=F,main="110mtj",
ylab="f(x)", xlab="aeg(s)",
col=gray(1:18/18) )
x=seq(10,35, length=500)
y=dnorm(x,mean=mean(X110mtj,      na.rm=T),      sd=sd(X110mtj,
na.rm=T))
lines(x,y,lwd=1.5)
qqnorm(X110mtj,main="110mtj"    ,    ylab="valimi    kvantiil",
xlab="normaaljaotuse kvantiil")
qqline(X110mtj)

#kettaheite tulemuste jaotus.
hist(ketas,breaks=seq(min(ketas,na.rm=T),max(ketas,na.rm=T),le
ngth=20),freq=F,main="kettaheide",
ylab="f(x)", xlab="kaugus(m)",
col=gray(1:19/19) )
x=seq(10,40, length=500)
```

```
y=dnorm(x,mean=mean(ketas, na.rm=T), sd=sd(ketas, na.rm=T))
lines(x,y,lwd=1.5)
qqnorm(ketas,main="kettaheide"   ,   ylab="valimi   kvantiil",
xlab="normaaljaotuse kvantiil")
qqline(ketas)

#teivashüppe tulemuste jaotus.
hist(teivas,breaks=seq(min(teivas,na.rm=T),max(teivas,na.rm=T)
,length=10),freq=F,main="teivashüpe",
ylab="f(x)", xlab="kõrgus(m)",
col=gray(1:11/11) )
x=seq(0,3, length=500)
y=dnorm(x,mean=mean(teivas, na.rm=T), sd=sd(teivas, na.rm=T))
lines(x,y,lwd=1.5)
qqnorm(teivas,main="teivashüpe",    ylab="valimi    kvantiil",
xlab="normaaljaotuse kvantiil" )
qqline(teivas)

#Joonised  kuue kaupa
par(mfrow=c(3,2))

#odaviske tulemuste jaotus.
hist(oda,
breaks=seq(10,max(oda,na.rm=T),length=17),freq=F,main="odavise
",
ylab="f(x)", xlab="kaugus(m)",
col=gray(1:17/17) )
x=seq(0,50, length=500)
y=dnorm(x,mean=mean(oda, na.rm=T), sd=sd(oda, na.rm=T))
lines(x,y,lwd=1.5)
qqnorm(oda,main="odavise"     ,     ylab="valimi     kvantiil",
xlab="normaaljaotuse kvantiil")
qqline(oda)

#1500m tulemuste jaotus.
hist(X1500m,freq=F,main="1500m",
ylab="f(x)", xlab="aeg(s)",
col=gray(1:11/11))
x=seq(0,1200, length=50)
y=dnorm(x,mean=mean(X1500m, na.rm=T), sd=sd(X1500m, na.rm=T))
lines(x,y,lwd=1.5)
qqnorm(X1500m,main="1500m",       ylab="valimi       kvantiil",
xlab="normaaljaotuse kvantiil" )
qqline(X1500m)

#1500m tulemuste jaotus ilma aeglaseima jooksuta.
a10v1=a10v[-160,]
hist(a10v1$X1500m,xlim=range(min(a10v1$X1500m,na.rm=T),680),
breaks=20,freq=F,main="1500m ilma aeglaseima jooksuta",
ylab="f(x)", xlab="aeg(s)",
col=gray(1:20/20) )
```

```
x=seq(200,800,length=500)
y=dnorm(x,mean=mean(a10v1$X1500m,                    na.rm=T),
sd=sd(a10v1$X1500m, na.rm=T))
lines(x,y,lwd=1.5)
qqnorm(a10v1$X1500m,main="1500m  ilma  aeglaseima  jooksuta"  ,
ylab="valimi kvantiil", xlab="normaaljaotuse kvantiil")
qqline(a10v1$X1500m)

#Andmestiku,  mis  koosneb  ainult  täielikulst  sooritatud
võistlustest, genereerimine
a10v2=a10v[!is.na(X100m)&!is.na(kaugus)&!is.na(kuul)&!is.na(kõ
rgus)&!is.na(X400m)
&!is.na(X110mtj)&!is.na(ketas)&!is.na(teivas)&!is.na(oda)&!is.
na(X1500m),]

#Täielikult sooritatud võistluste arv
length(a10v2$nimi)

#Joonised  kahe kaupa
par(mfrow=c(3,2))

#kogusumma jaotus
hist(a10v2$kokku,freq=F,breaks=seq(min(a10v2$kokku,na.rm=T),ma
x(a10v2$kokku,na.rm=T),length=18),main="kogusummad, n=139",
ylab="f(x)", xlab="punktid",
col=gray(1:18/18))
x=seq(1000,6000, length=50)
y=dnorm(x,mean=mean(a10v2$kokku,  na.rm=T),  sd=sd(a10v2$kokku,
na.rm=T))
lines(x,y,lwd=1.5)
qqnorm(a10v2$kokku,main="kogusummad,    n=139",    ylab="valimi
kvantiil", xlab="normaaljaotuse kvantiil" )
qqline(a10v2$kokku)

#Andmestiku jagamine kaheks 3350p pealt
a10v3=a10v2[a10v2$kokku>=3350,]
length(a10v3$nimi)
a10v4=a10v2[a10v2$kokku<3350,]
length(a10v4$nimi)

#kogusumma jaotus (>=3500)
hist(a10v3$kokku,freq=F,breaks=seq(min(a10v3$kokku,na.rm=T),ma
x(a10v3$kokku,na.rm=T),length=18),main="kogusummad      >=3350,
n=49",
ylab="f(x)", xlab="punktid",
col=gray(1:18/18))
x=seq(1000,6000, length=50)
y=dnorm(x,mean=mean(a10v3$kokku,  na.rm=T),  sd=sd(a10v3$kokku,
na.rm=T))
lines(x,y,lwd=1.5)
```

```
qqnorm(a10v3$kokku,main="kogusummad       >=3350,       n=49",
ylab="valimi kvantiil", xlab="normaaljaotuse kvantiil" )
qqline(a10v3$kokku)

#kogusumma jaotus (<3500)
hist(a10v4$kokku,freq=F,breaks=seq(min(a10v4$kokku,na.rm=T),ma
x(a10v4$kokku,na.rm=T),length=18),main="kogusummad       <3350,
n=90",
ylab="f(x)", xlab="punktid",
col=gray(1:18/18))
x=seq(1000,6000, length=50)
y=dnorm(x,mean=mean(a10v4$kokku,  na.rm=T),  sd=sd(a10v4$kokku,
na.rm=T))
lines(x,y,lwd=1.5)
qqnorm(a10v4$kokku,main="kogusummad <3350, n=90", ylab="valimi
kvantiil", xlab="normaaljaotuse kvantiil" )
qqline(a10v4$kokku)

#Joonised  kahe kaupa
par(mfrow=c(1,2))

#Täieliku võistluse teinud sportlaste arv
keskmised=by(a10v2$kokku, a10v2$nimi,mean)
length(table(keskmised))

#keskmiste punkti summade jaotus
hist(keskmised,freq=F,breaks=seq(min(keskmised,na.rm=T),max(ke
skmised,na.rm=T),length=18),main="keskmised       kogusummad,\n
n=61",
ylab="f(x)", xlab="punktid",
col=gray(1:18/18))
x=seq(1000,6000, length=50)
y=dnorm(x,mean=mean(keskmised,    na.rm=T),    sd=sd(keskmised,
na.rm=T))
lines(x,y,lwd=1.5)
qqnorm(keskmised,main="keskmised      kogusummad,\n      n=61",
ylab="valimi kvantiil", xlab="normaaljaotuse kvantiil" )
qqline(keskmised)

#Mudelid lõpptulemuse ennustamiseks
detach(a10v)
attach(a10v2)

# Erinevate punktide ja tulemusete korreleerumine kogusummaga
cor(a10v2[,2:22],a10v2[,22])

#Esimesed mudelid
#kaugus
mudel=lm(kokku~kaugus_p)
mudel
summary(mudel)
```

```
#400m
mudel=lm(kokku~X400m_p)
mudel
summary(mudel)

#Joonised  ühe kaupa
par(mfrow=c(1,1))
#Täiustatud mudel
mudel=lm(kokku~X400m_p+kaugus_p)
mudel
summary(mudel)
plot(((X400m_p*2.396)/(2.396+4.582)+(kaugus_p*4.582)/(2.396+4.
582)),kokku,main="Kogusumma  prognoosimine  \n  kaugushüppe  ja
400m punktid",
ylab="kogusumma  punktid",  xlab="punktid  \n  (400m  jooksu
punktid*2.396)/(2.396+4.582)+(kaugushüppe
punktid*4.582)/(2.396+4.582)")
prognoos=predict(mudel,
data.frame(kaugus_p=c(50,400,600),X400m_p=c(50,400,600) ))
lines(c(50,400,600),prognoos)
prognoos=predict(mudel,
data.frame(kaugus_p=c(50:600),X400m_p=c(50:600)),interval="pre
diction")
lines(50:600, prognoos[,2],lty=2)
lines(50:600, prognoos[,3],lty=2)
prognoos=predict(mudel,
data.frame(kaugus_p=c(50:600),X400m_p=c(50:600)),interval="con
fidence")
lines(50:600, prognoos[,2],lty=2)
lines(50:600, prognoos[,3],lty=2)

#Mudeli jääkide iseloomustamine
qqnorm(resid(mudel))
qqline(resid(mudel))

#Mudeli jääkide hajuvuse iseloomustamine
plot(((X400m_p*2.396)/(2.396+4.582)+(kaugus_p*4.582)/(2.396+4.
582)), resid(mudel))
```