Doping

Risto Hinno

Saturday, June 27, 2015

Sissejuhatus

Sander "Gamma" Lognormaalsele meeldib võistelda. Eelmine nädal käis ta rahvusvahelisel statistikaolümpiaadil, kus ta ei suutnud enamat pronksmedalist. Olümpiaadil tuli lahendada mitmeid huvitavaid statistikaülesandeid. Teoreetilises osas oli vaja tuvastada, kas mõni hinnang on nihkega või nihketa. Praktiline osa algas aja peale andmete sisestamisega SASi. Seekord oli praktiline osa eriti põnev, kuna töötati reaalsete andmetega: andmetabelis oli 11 vaatlust ning 2 tunnust. Sellele järgnes osa "Märka olulisi tunnuseid", kus tuli aja peale SASi väljundist üles leida tunnused, mille p-väärtus on väiksem kui 0.05.

Oma tulemusi analüüsides sai Sandrile selgeks, et konkurentidele jäi ta alla viimases, väikeste p-väärtuste märkamise voorus. Ta kahtlustas, et konkurentide paremus tulenes tähelepanu tõstvate ainete manustamisest ning statistikaolümpiaadil tuleks kehtestada dopingutestid. Aga kuidas seda teha? Õnneks leidis ta, et Krista Fischer on tegelenud ühe dopingujuhtumi analüüsimisega, ning on koos Donald A. Berryga kirjutanud artikli Statisticians Introduce Science to International Doping Agency: The Andrus Veerpalu Case.

Selles praktikumis uurimegi sellesama dopingujuhtumi näitel, kuidas kasvuhormooni dopingutesti piirmäärad seati ning mille vastu eksiti.

Kasvuhormoonist

Kasvurhormoon on inimkehas toodetud aine, mille ülesanne on reguleerida kehapikkust, lihaste ja organite kasvu. Ravimina on kasutusel sünteetiline kasvuhormoon, mida kasutatakse näiteks laste kasvudefektide ravis ja aidsihaigete üldseisundi parandamiseks. Kuigi teaduslikult pole tõestatud, et sünteetilise kasvuhormooni pruukimine tõstab sportlikke tulemusi (vt Stanfordi ülikooli teadlaste meta-analüüs, mis võtab kokku 27 uuringu tulemused), on spordiringkond siiski arvamusel selle positiivsest mõjust ning 1989. aastal lisas Rahvusvaheline Olümpiakomitee kasvuhormooni keelatud ainete nimistusse.

Sünteetilise kasvuhormooni kasutamist on raske tuvastada. Üks põhjustest on kasvuhormooni kontsentratsiooni suur varieeruvus päeva lõikes ning pulsseeruv sekretsioon.

Saksa teadlaste 2009. aastal väljatöötatud test ei kontrolligi kasvuhormooni taset, vaid hormooni erinevate molekulitüüpide omavahelist tasakaalu. Eeldatakse, et erinevate isovormide suhe on ajas konstantne. Kuna arvatakse, et sünteetiline kasvuhormooni süstimisel muutub vaid ühe isovormi tase veres. Dopingutesti idee seisnebki testimises, kas kahe isovormi suhe on statistiliselt erinev loomulikust suhtest.

Andmestik

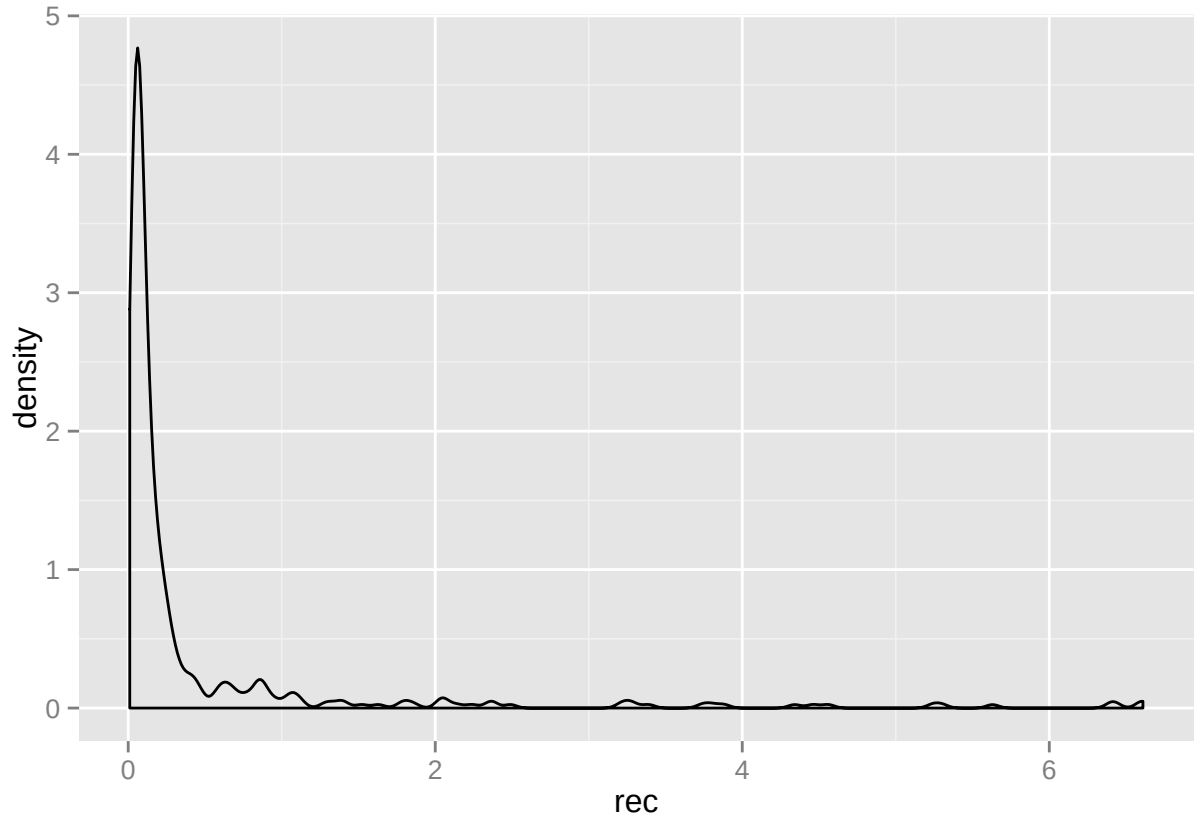
Laadi ÕISist alla andmestik doping.csv ja loe töökeskkonda. Andmestikus on järgmised tunnused:

- ethnicity - etnilisus: kas african või caucasian
- kit - dopingutesti erinevad variandid (kit1 ja kit2)
- rec - kasvuhormooni isovormi rec kontsentratsioon
- pit - kasvuhormooni isovormi pit kontsentratsioon

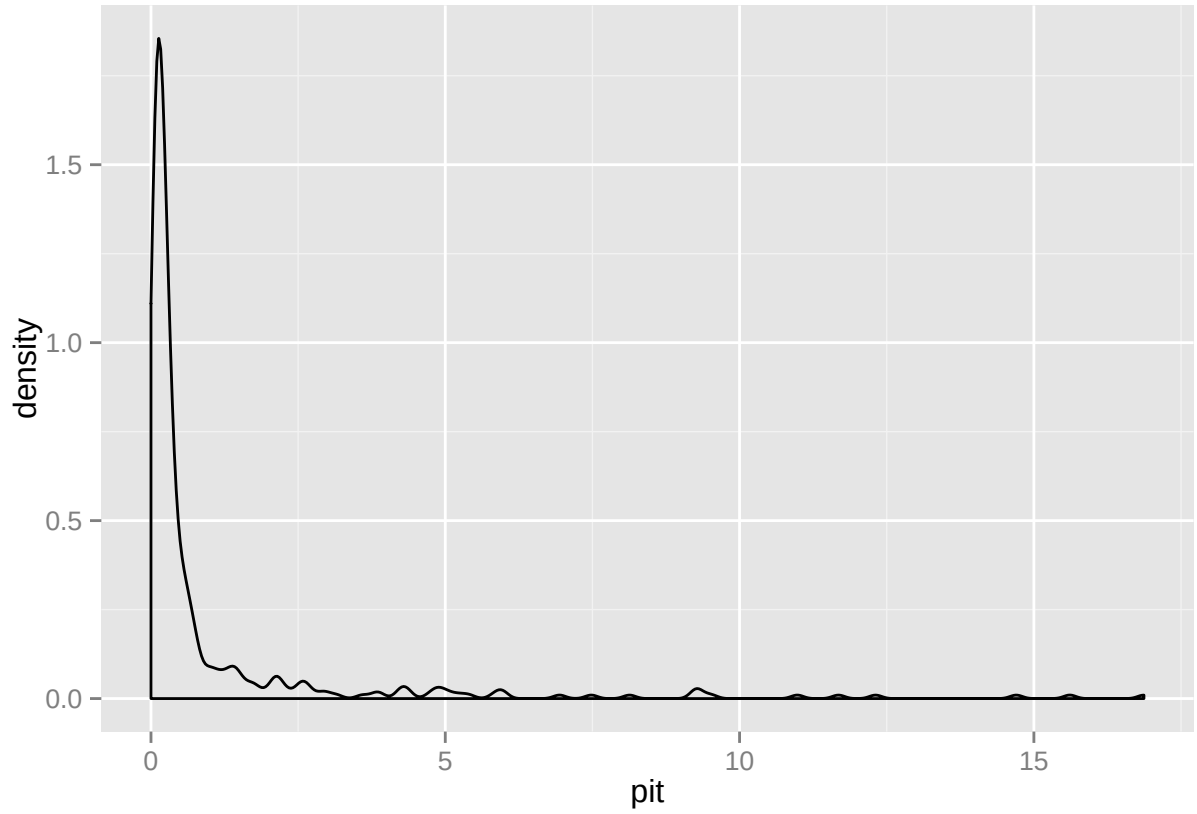
Ülesanne 1 (1 punkt) - andmetega tutvumine

Lisa andmetabelisse tunnus ratio, mis näitab tunnuste rec ja pit suhet. Visualiseeri tunnuste rec, pit ja ratio jaotusi.

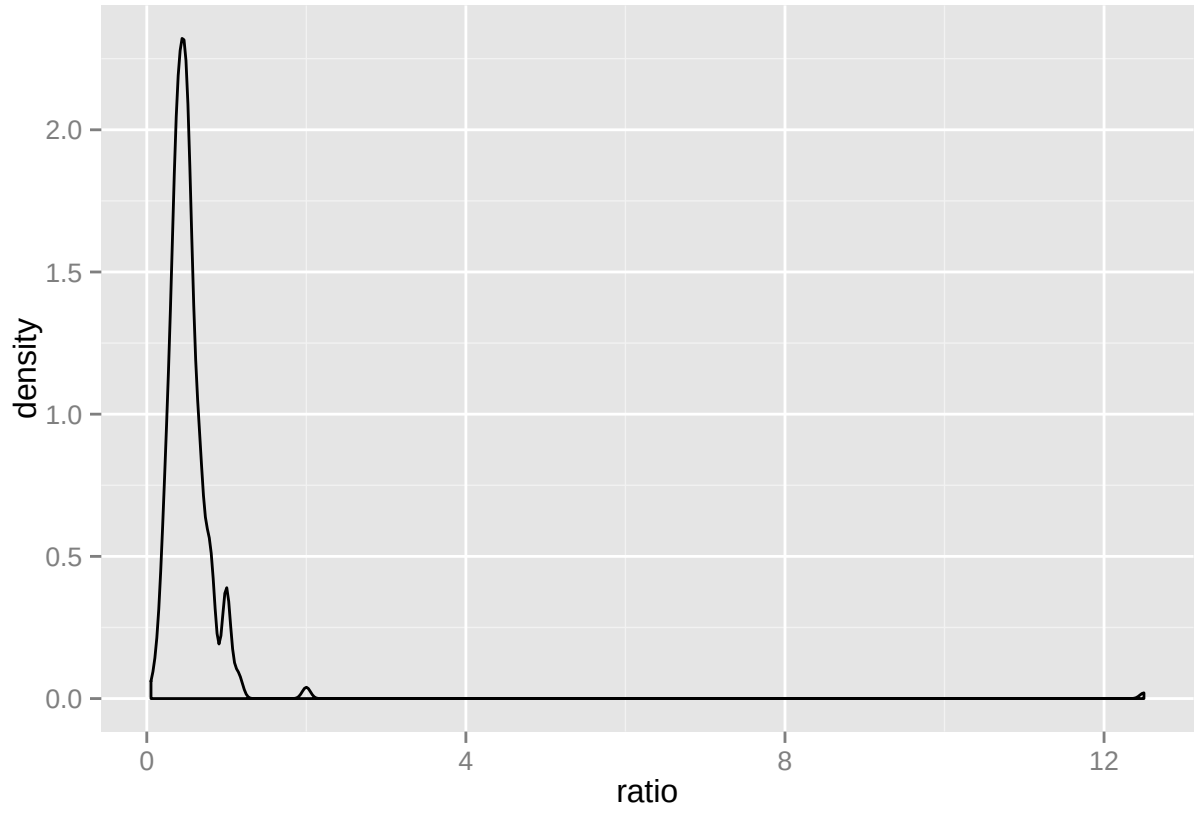
```
doping=read.csv("./data/doping.csv")
#lisame ratio
doping$ratio=doping$rec/doping$pit
#rec jaouts
library(ggplot2)
ggplot(doping, aes(x=rec))+
    geom_density()
```




```
#pit
ggplot(doping, aes(x=pit))+
    geom_density()
```




```
#ratio
ggplot(doping, aes(x=ratio))+
    geom_density()
```



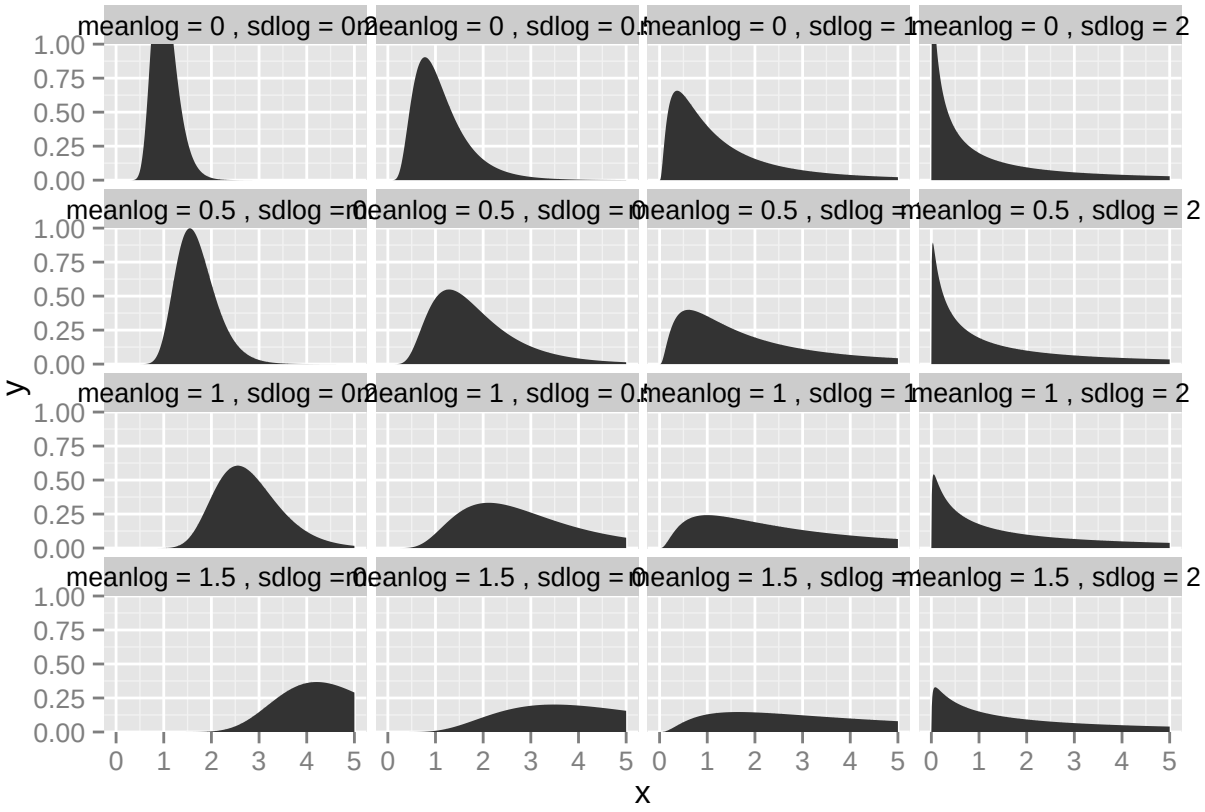
Ülesanne 2 (3 punkti) - jaotustest

Kodutööna lugesid artiklit Statisticians Introduce Science to International Doping Agency: The Andrus Veerpalu Case. Said teada, et dopingutest põhines kasvuhormooni isovormide suhtel ehk tunnusel ratio. Mingi hetk kasutati isovormide suhte modelleerimiseks log-normaalset jaotust, mingil hetkel see enam ei sobinud ning kasutusele võeti gammajaotus.

Aga millised näevad välja log-normaaljaotus ning gammajaotus? Selleks visualiseeri neid.

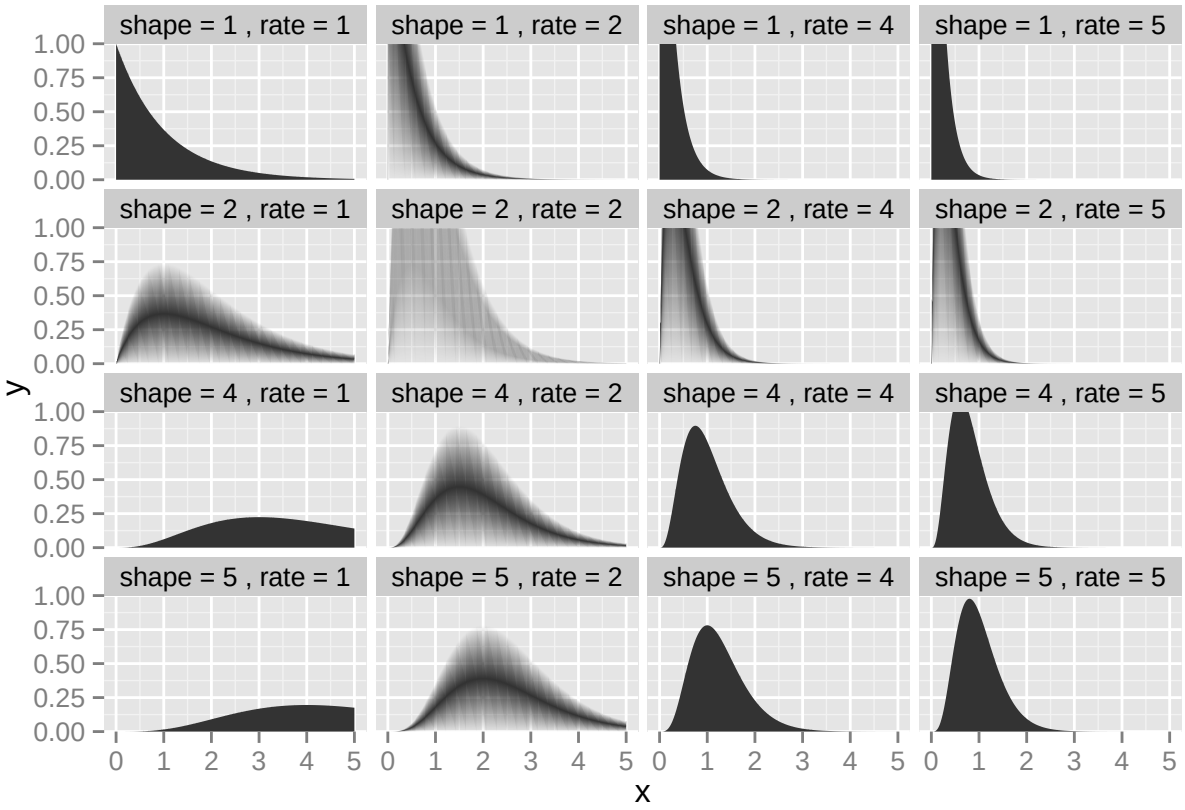
```
#log-normaalse jaoutse tihedus
library(ggplot2)
library(dplyr)

i = 1
df_list = list()
for(m in c(0, 0.5, 1, 1.5)){
    for(s in c(0.25, 0.5, 1, 2)){
        # tiheduse graafiku jaoks x ja y koordinaadid
        x = seq(0, 5, 0.01)
        y = dlnorm(x, meanlog = m, sdlog = s)

        df_list[[i]] = data.frame(x=x, y=y,
                                  group = paste("meanlog =", m, ", sdlog =", s))
        i = i + 1
    }
```

```
}
df = rbind_all(df_list)
ggplot(df, aes(x, y)) + geom_area() + facet_wrap(~ group) +
    coord_cartesian(ylim = c(0, 1))
```



Joonista ka gammajaotuse tihedusfunktsioonid erinevate alfa on{1,2,3,4,5} ja beeta on {1,2,3,4,5} jaoks. Selgita, kuidas muutub jaotus, kui muudame kumbagi parameetrit.

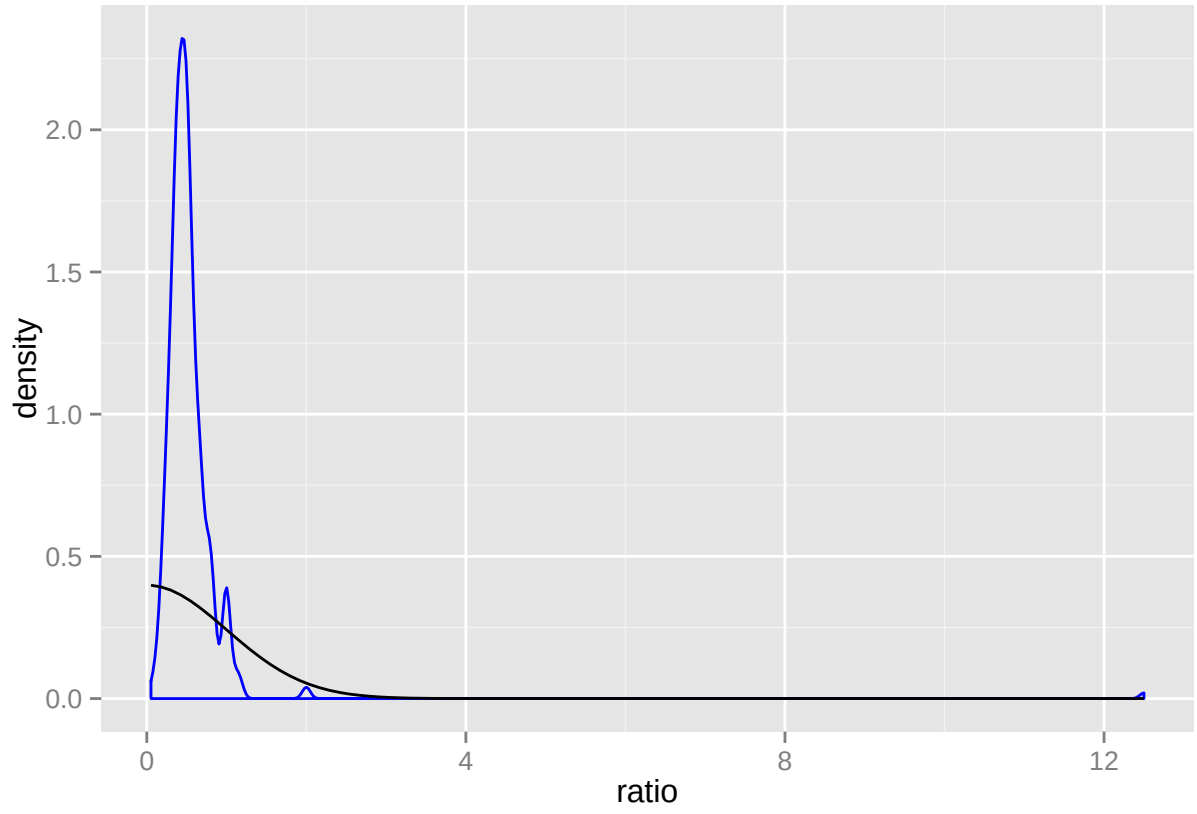
```
i = 1
df_list_gamma = list()
for(m in c(1, 2, 2, 4, 5)){
    for(s in c(1, 2, 2, 4, 5)){
        # tiheduse graafiku jaoks x ja y koordinaadid
        x = seq(0, 5, 0.01)
        y = dgamma(x, shape = m, rate = s)

        df_list_gamma[[i]] = data.frame(x=x, y=y,
                                        group = paste("shape =", m, ", rate =", s))
        i = i + 1
    }
}
df_gamma = rbind_all(df_list_gamma)
ggplot(df_gamma, aes(x, y)) + geom_area() + facet_wrap(~ group) +
    coord_cartesian(ylim = c(0, 1))
```



Tihti eeldatakse statistikas, et tunnus on normaaljaotusega. Kas isovormide suhe võiks põhimõtteliselt olla normaaljaotusega? Põhjenda.

Ei. Üritan sobitada peale normaaljaotust joonisel.

```
ggplot(doping, aes(x=ratio))+
    geom_density(colour="blue")+
    stat_function(fun=dnorm)
```



WADA metoodika analüüs

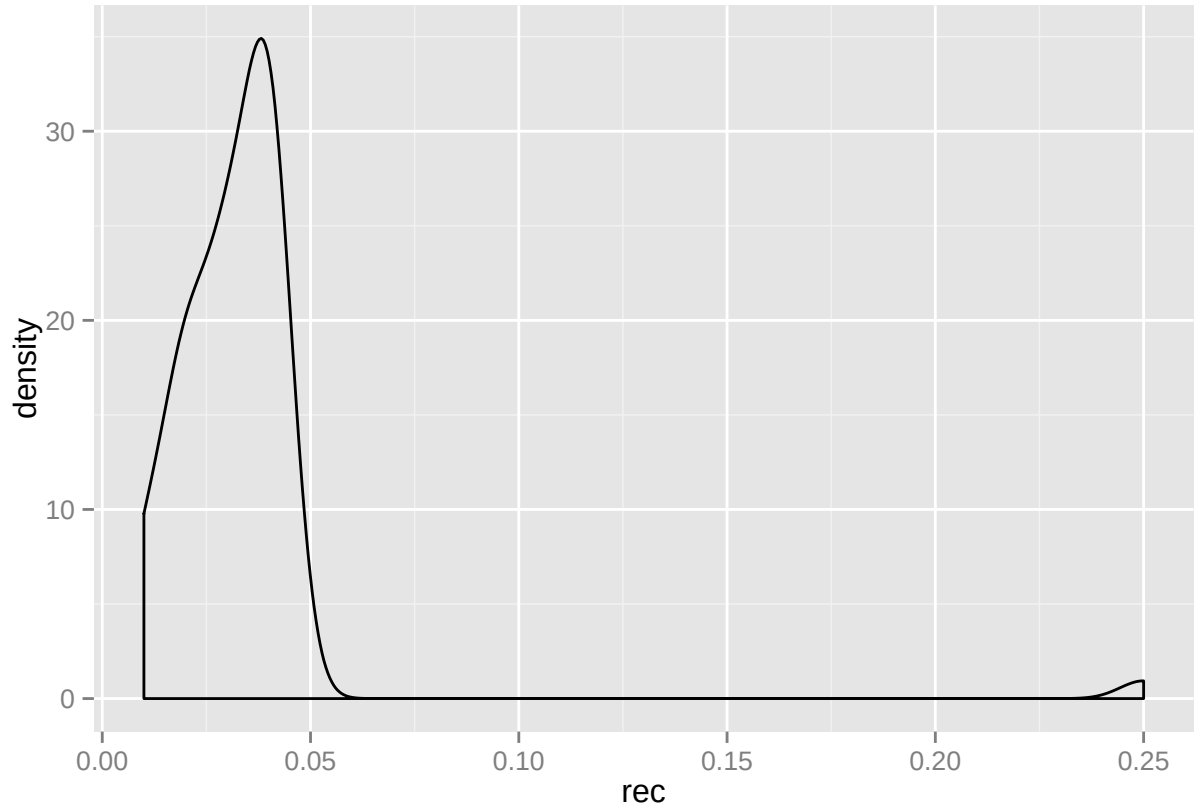
Lühikokkuvõte, kuidas määras otsustuspiirid WADA.

- Analüüsist jäeti välja andmepunktid, kus rec või pit kontsentratsioon oli väiksem kui 0.05.
- Neljale osagrupile (kit1 - valged, kit1 - mustanahalised, kit2 - valged, kit2 - mustanahalised) sobitati parameetriline jaotus.
- WADA väitel sobitus lognormaalne jaotus andmetele kõige paremini.
- Otsustuspiiriks võeti 99.99% log-normaaljaotuse kvantiil.
- Võeti kasutusele mustanahaliste piirmäärad, sest need olid suuremad.

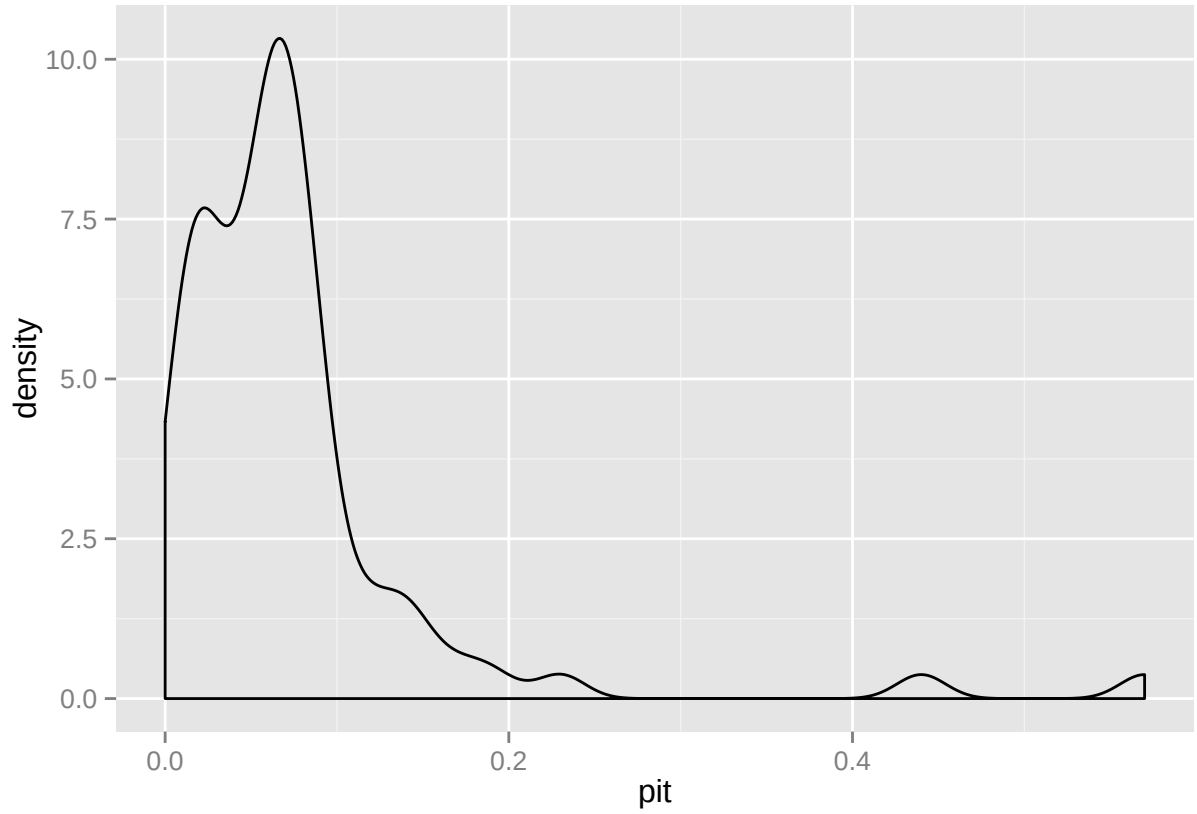
Ülesanne 3 (2 punkti) - andmete filtreerimine

- Mis sa arvad, miks jäeti analüüsist välja andmepunktid, kus rec või pit kontsentratsioon oli väiksem kui 0.05? Ei tea täpselt.Kuna nende ratio oli kõikuv?
- Visualiseeri hajuvusdiagrammi abil, millised andmepunktid jäid analüüsist välja.

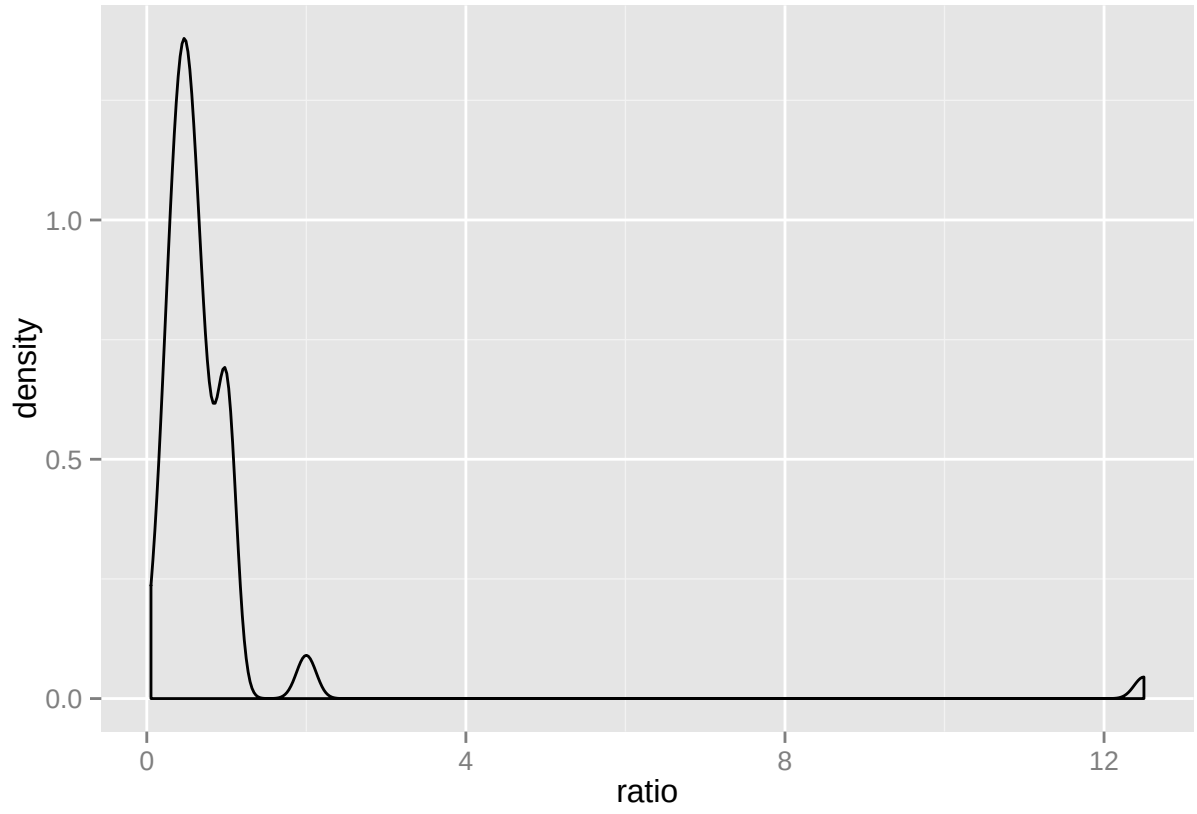
```
valjajaanud=subset(doping, rec<0.05| pit<0.05)

ggplot(valjajaanud, aes(x=rec))+
    geom_density()
```




```
ggplot(valjajaanud, aes(x=pit))+
    geom_density()
```




```
ggplot(valjajaanud, aes(x=ratio))+
    geom_density()
```



Edasises kasuta andmestikku, kus rec ja pit on suuremad kui 0.05.

```
doping_subset=subset(doping, rec>0.05& pit>0.05)
```

Ülesanne 4 (4 punkti) - parameetrilise jaotuse sobitamine

Log-normaaljaotusel on kaks parameetrit. Kuidas leiad sellised parameetrite väärtused, mille korral jaotus sobiks andmetega kõige paremini? Leia need parameetrid "kit1 - valged" osapopulatsiooni jaoks.

```
library(MASS)
log_norm_param=fitdistr(x=doping_subset[c(doping_subset$kit==1&doping_subset$ethnicity=="caucasian"), c
         densfun="log-normal")

meanlog1=log_norm_param$estimate[1]
sdlog1=log_norm_param$estimate[2]
```

Ka gammajaotusel on kaks parameetrit. Kuidas leiad sellised parameetrite väärtused, mille korral jaotus sobiks andmetega kõige paremini? Leia need parameetrid "kit1 - valged" osapopulatsiooni jaoks.

```
gamma_param=fitdistr(x=doping_subset[c(doping_subset$kit==1&doping_subset$ethnicity=="caucasian"), c("ra
                        densfun="gamma")

shape1=gamma_param$estimate[1]
rate1=gamma_param$estimate[2]
```

Leia kummagi jaotuse 99.99% kvantiil.

```
#log-normal
qlnorm(0.9999, meanlog = meanlog1, sdlog = sdlog1)
```

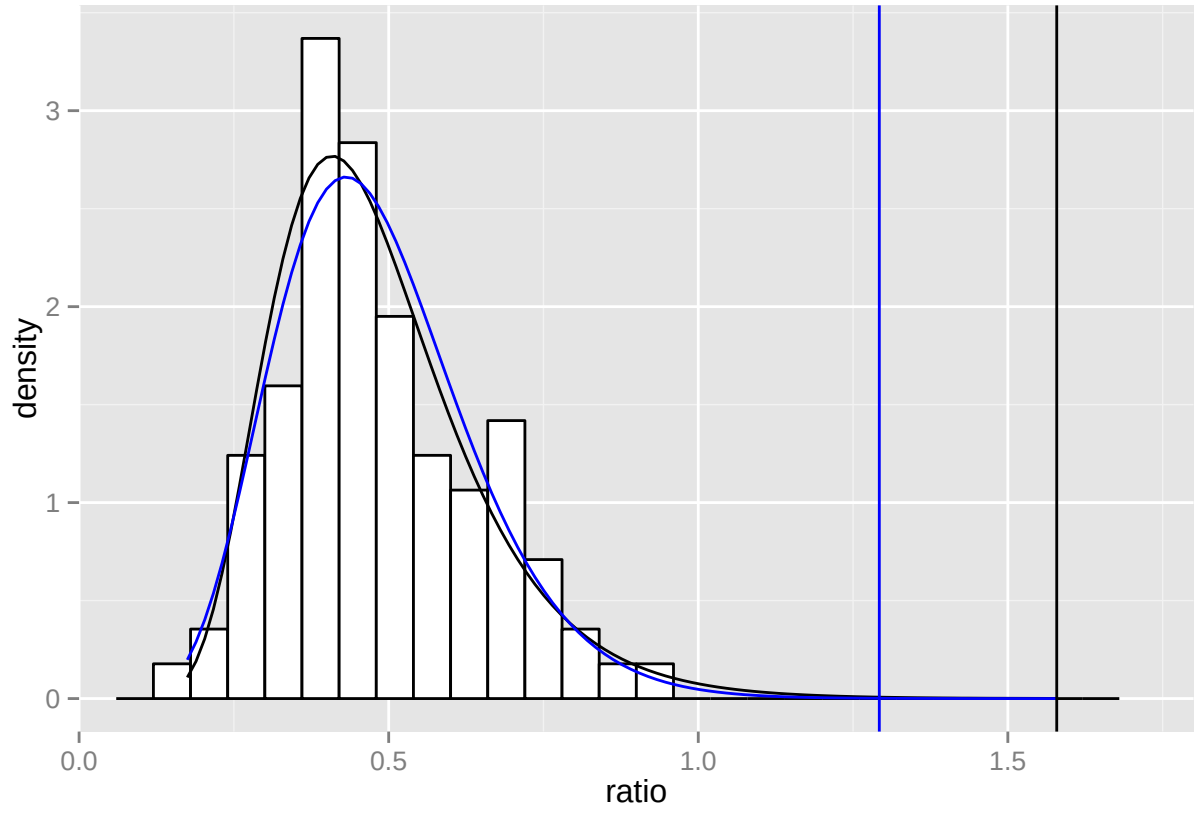
```
## [1] 1.578964
```

```
#gamma
qgamma(0.9999, shape=shape1, rate = rate1)
```

```
## [1] 1.292397
```

Visualiseeri ühel joonisel koos andmetega nii sobitatud log-normaal kui ka gammajaotust. Lisa joonisele 99.99% kvantiil.

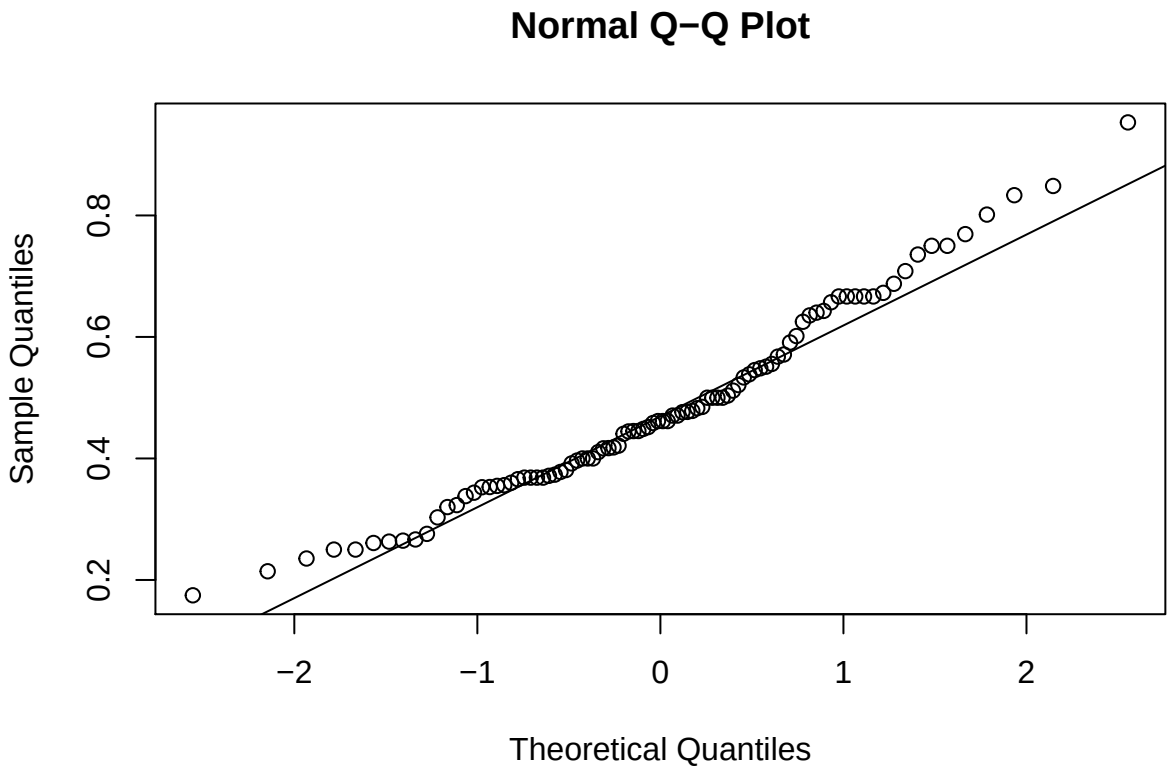
```
kit1_caucasian=doping_subset[c(doping_subset$kit==1&doping_subset$ethnicity=="caucasian"),]

ggplot(kit1_caucasian, aes(x=ratio))+
    geom_histogram(aes(y=..density..), binwidth=0.06, fill="white", colour="black" )+
    stat_function(fun=dlnorm, args = list(meanlog=meanlog1, sdlog=sdlog1))+
    stat_function(fun=dgamma, args=list(shape=shape1, rate=rate1), colour="blue")+
    geom_vline(xintercept=qlnorm(p=0.9999, meanlog = meanlog1, sdlog = sdlog1))+
    geom_vline(xintercept=qgamma(p=0.9999, shape = shape1, rate = rate1), colour="blue")+
    coord_cartesian(xlim=c(0, 1.8))
```



Ülesanne 5 (3 punkti) - kas jaotus sobib andmetega

- Kas eelnevalt sobitatud log-normaaljaotus võiks sobida andmetega? Mille alusel otsustad?
- Aga kas gammajaotus võiks sobida andmetega?
- Praktikumis arutasime märksõnu QQplot ja Kolmogorov-Smirnovi test.
- Oletame, et mõlemad jaotused sobisid andmetega. Mille põhjal langetad otsuse, kumb sobib paremini?
- WADA väitis, et log-normaaljaotus sobib andmetele kõige paremini. Kas said sama tulemuse?

```
qqnorm(kit1_caucasian$ratio)
qqline(kit1_caucasian$ratio)
```




```
#gamma sobivuse test
ks.test(kit1_caucasian$ratio,"pgamma", shape=shape1, rate=rate1)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  kit1_caucasian$ratio
## D = 0.053, p-value = 0.9541
## alternative hypothesis: two-sided
```

```r
#log-normali sobivuse test
ks.test(kit1_caucasian$ratio,"dlnorm", meanlog=meanlog1, sdlog=sdlog1)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  kit1_caucasian$ratio
## D = 2.4381, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Ei saanud. Sain, et gammajaotus sobib. Log-normaalsele jaotusele ei allu statistiliselt olulisel määral. Kui mõlemad sobiks, siis üks võimalus oleks vaadata p-väärtust.

Ülesanne 6 (2 punkti) - piirmäärad alamgruppide kaupa

Sobita nüüd kõigile neljale alamgrupile log-normaaljaotus ja leia selle 99.99% kvantiil.

```r
kit2_caucasian=doping_subset[c(doping_subset$kit==2&doping_subset$ethnicity=="caucasian"),]
kit1_african=doping_subset[c(doping_subset$kit==1&doping_subset$ethnicity=="african"),]
kit2_african=doping_subset[c(doping_subset$kit==2&doping_subset$ethnicity=="african"),]
#valged kit2
log_norm_param2=fitdistr(x=kit2_caucasian$ratio, densfun="log-normal")
meanlog2=log_norm_param2$estimate[1]
sdlog2=log_norm_param2$estimate[2]
qlnorm(0.9999, meanlog = meanlog2, sdlog = sdlog2)
```

```
## [1] 1.83209
```

```r
#mustad kit1
log_norm_param3=fitdistr(x=kit1_african$ratio, densfun="log-normal")
meanlog3=log_norm_param3$estimate[1]
sdlog3=log_norm_param3$estimate[2]
qlnorm(0.9999, meanlog = meanlog3, sdlog = sdlog3)
```

```
## [1] 2.252636
```

```r
#mustad kit2
log_norm_param4=fitdistr(x=kit2_african$ratio, densfun="log-normal")
meanlog4=log_norm_param4$estimate[1]
sdlog4=log_norm_param4$estimate[2]
qlnorm(0.9999, meanlog = meanlog4, sdlog = sdlog4)
```

```
## [1] 2.004172
```

```r
#kittide lõikes, kit1
kit1=doping_subset[c(doping_subset$kit==1),]
log_norm_param5=fitdistr(x=kit1$ratio, densfun="log-normal")
meanlog5=log_norm_param5$estimate[1]
sdlog5=log_norm_param5$estimate[2]
qlnorm(0.9999, meanlog = meanlog5, sdlog = sdlog5)
```

```
## [1] 1.884795
```

```
#kit2
kit2=doping_subset[c(doping_subset$kit==2),]
log_norm_param6=fitdistr(x=kit2$ratio, densfun="log-normal")
meanlog6=log_norm_param6$estimate[1]
sdlog6=log_norm_param6$estimate[2]
qlnorm(0.9999, meanlog = meanlog6, sdlog = sdlog6)
```

```
## [1] 1.920686
```

Ametlik piirmäär kit1 korral oli 1.81 ja kit2 korral 1.68. Kas said sarnased tulemused? Kit1 korral enam-vähem, kit2 korral on erinevus suurem.

Kuidas verifitseeris otsustuspiire WADA?

Esmased otsustuspiirid on määratud. Nüüd tuleb neid verifitseerida. Kuidas tegi seda WADA?

- Rutiinsete dopingukontrollide käigus koguti aastatel 2009-2011 kit1 kohta 3547 mõõtmist ja kit2 kohta 617 mõõtmist.
- Nendes andmetes puudub tunnus ethnicity.
- Nüüd filtreeriti välja andmepunktid, kus rec kontsentratsioon oli väiksem kui 0.1 ja pit kontsentratsioon oli väiksem kui 0.05.
- Lognormaaljaotus ei sobinud. Kasutati gammajaotust.
- Visati välja 10 imelikku (liiga kõrget) andmepunkti.
- Selle andmestiku põhjal arvutatud kvantiilid tulid väiksemad kui esmase uuringu kvantiilid. Järeldati, et esmased piirmäärad on verifitseeritud.

Laadi ÕISist alla andmestik doping_verification.csv.

```
verification=read.csv("./data/doping_verification.csv.")
```

Boonusülesanne (kuni 5 punkti)

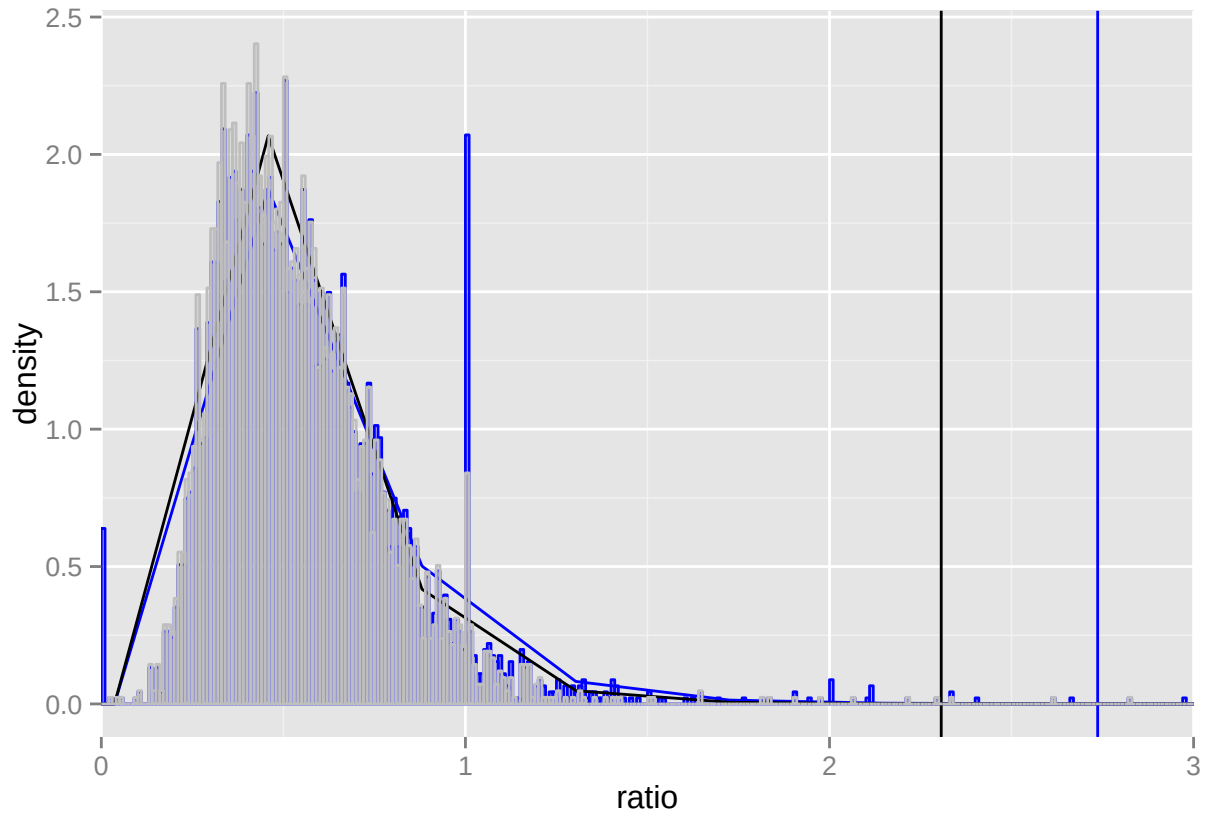
Uuri ise midagi põnevat. Näiteks testi, kas tõesti enam log-normaaljaotus ei sobi, või uuri, milline mõju oli imelike andmepunktide väljaviskamisel.

```
#arvutame ratio
verification$ratio=verification$rec/verification$pit
#pean ka miinused ja nullid välja võtma, muidu ei tööta!!!!
param_verif1=fitdistr(x=(verification$ratio[!is.infinite(verification$ratio)&!is.na(verification$ratio)&

meanlog_verif1=param_verif1$estimate[1]
sdlog_verif1=param_verif1$estimate[2]

#viskame välja sodi
verif_subset=subset(verification, rec>=0.01 & pit>=0.05)
```

```
param_verif2=fitdistr(x=verif_subset$ratio, densfun="log-normal")

meanlog_verif2=param_verif2$estimate[1]
sdlog_verif2=param_verif2$estimate[2]

ks.test(verif_subset$ratio,"dlnorm", meanlog=meanlog_verif2, sdlog=sdlog_verif2)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  verif_subset$ratio
## D = 1.8102, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
#mis juhtus, kui viskasime välja osad vaatused
    ggplot(verification, aes(ratio))+
        geom_histogram(aes(y=..density..), binwidth=0.01, fill="blue", colour="blue", alpha=0.1 )+
        geom_vline(xintercept=qlnorm(p=0.9999, meanlog = meanlog_verif1, sdlog = sdlog_verif1), colour=
        geom_vline(xintercept=qlnorm(p=0.9999, meanlog = meanlog_verif2, sdlog = sdlog_verif2))+
        stat_function(fun=dlnorm, args = list(meanlog=meanlog_verif1, sdlog=sdlog_verif1), colour="blue
        stat_function(fun=dlnorm, args = list(meanlog=meanlog_verif2, sdlog=sdlog_verif2))+
        coord_cartesian(xlim=c(0, 3))+
        geom_histogram(data=verif_subset,aes(y=..density..), binwidth=0.01, fill="grey", colour="grey",
```



Sinine on see osa histogrammis, mida pärast väljaviskamist alles ei jäänud.

Testi kriitika

See, kas tegu on gamma- või lognormaaljaotusega on tegelikult mõnes mõttes pseudoprobleem. Tuleb silmas pidada, et me tahame hinnata väga äärmuslikku, 99,99% kvantiili. Kui me valimi põhjal mingi standardse testiga kontrollime parameetrilise jaotuse kehtivust ja jääme nullhüpoteesi juurde, siis see tähendab seda, et suurem osa andmete jaotusest, nn jaotuse "keha" sobib selle jaotuse mudeliga. See test ei ütle midagi jaotuse "saba" kaugema otsa kohta. Ei ole suurt mõtet valideerimisuuringutes näha vaeva sellega, kas algul eeldatud jaotus paika peab. Pigem tuleks näha vaeva sellega, et uurida otsusepiirist üle minevate tulemuste kohta mingitki tausta - kas on vähimatki lisatõendust dopingutarvitamise kohta? Ega ikka muudmoodi ei saagi seda valepositiivse tulemuse tõenäosust kätte. Ja kui see pole võimalik. . . siis kas saabki kehaomase aine lisadoseerimist täie kindlusega testida?

Ülesanne 7 (5 punkti) - bootstrap usaldusintervall otsustuspiirile

Testi otsustuspiirid olid määratud vaid 106 sportlase põhjal, kui testiga hakati juba sportlaseid "vahele võtma". Samas eeldati, et test teeb vaid 1 vea 10000 testis. Sellise täpsuse saamiseks oli valimimaht ilmselgelt liiga väike.

Artiklis oli kirjas:

WADA scientists claim that the test with the resulting DL (decision limit) has a specificity of at least 99.99%. The claimed false-positive rate of less than 1 in 10,000 is quite remarkable from a sample size of less than 200! Clearly, it relies strongly on the parametric form of testing results.

Leia bootstrap 95% usaldusintervall dopingutesti piirmäärale (ehk 99.99% kvantiilile). Visualiseeri saadud tulemust.

Bootstrap on üks viis leida valimihinnangule (näiteks meid huvitavale kvantiilile) usaldusintervall. Kui bootstrap on sinu jaoks uus, on abiks järgnev Coursera kursuse Data Analysis and Statistical Inference video Unit 4 Part 2 - Bootstrapping. Juhul, kui sa ei soovi sellele kursusele registreeruda, saad seda videot vaadata ka siit lingilt

```
piir=numeric()

for (i in 1:10000) {
    param=fitdistr(x=sample(verif_subset$ratio, replace=T), densfun="log-normal")
    meanlog=param$estimate[1]
    sdlog=param$estimate[2]
    piir[i]=qlnorm(0.9999, meanlog = meanlog, sdlog = sdlog)
}

#piirid 95% jagunemise järgi
quantile(piir, c(.025, .5, .975))
```

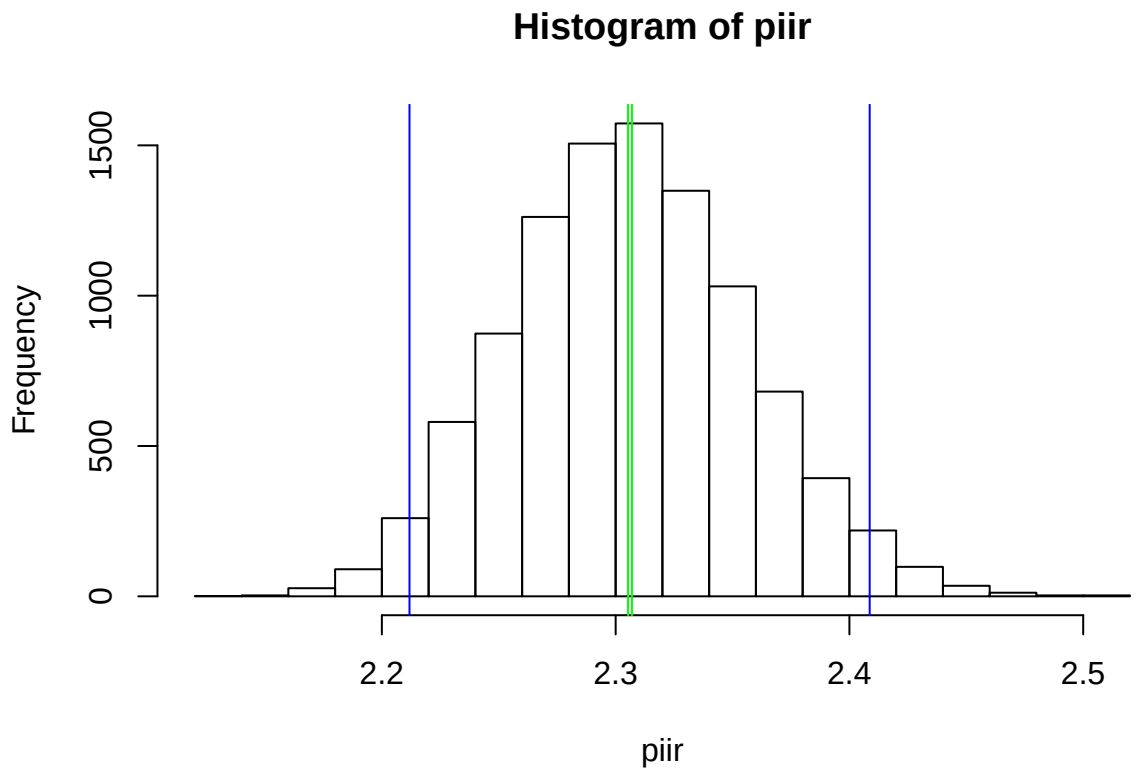
```
##     2.5%      50%    97.5%
## 2.211816 2.304809 2.408670
```

```
#teine variant
se=sd(piir)/sqrt(length(piir))
keskmine=mean(piir)
keskmine+c(-1,1)*1.65*se
```

```
## [1] 2.305304 2.306973
```

Plotime tulemused

```
hist(piir)
#piirid 95% reaalsete andmete jagunemise jagunemise järgi
abline(v=quantile(piir, c(.025, .5, .975))[1], col="blue")
abline(v=quantile(piir, c(.025, .5, .975))[3], col="blue")

#teine variant, kus 95% t-jaotuse järgi
abline(v=keskmine+c(-1)*1.65*se, col="green")
abline(v=keskmine+c(1)*1.65*se, col="green")
```



Teine variant on kitsam, kuna tegin mitu iteratsiooni (10 000) ja standard error on selle tõttu väike. Seega reaalne andmete jaotus (variant 1) võib anda täpsema tulemuse.

Ülesanne 8 (3 punkti) - valepositiivse testitulemuse tõenäosus

CASis kohal käinud statistikud (1 Eestist, 1 vägagi nimekas professor USAst) tegid selgeks, et kui test on positiivne, siis dopingutarvitamise tõenäosus jääb vahemikku 10-90%. 90% on õige siis, kui WADA poolt väidetav spetsiifilisus 99,99% (ehk siis valepositiivse tõenäosus 0,01%) kehtib. 10% on õige siis, kui see on tegelikult 99,9%. WADA andmed ei võimalda kindlaks teha, kas ta tegelikult on 99,99% või 99,9%. Sellega oli nõus ka CAS

Kuidas leida dopingutarvitamise tõenäosust, kui sportlane sai positiivse testitulemuse?

Näpunäited:

- Oletame, et dopingutarvitajaid on 2%, testi tundlikkus on 25%, spetsiifilisus on 99.9% ning oleme testinud 10000 sportlast. Täida nende eelduste põhjal järgmine tabel:

Meeldetuletus:

Testi tundlikkus=õiged positiivsedõiged positiivsed + valenegatiivsed

Testi spetsiifilisus=õiged negatiivsedõiged negatiivsed + valepositiivsed

Mis on tõenäosus, et positiivse dopingutesti korral on sportlane ka tegelikult dopingut tarvitanud?

```
tabel=read.csv("./data/tundlikkus.csv")
tabel
```

```
##                         X Positiivne.test Negatiivne.test KOKKU
## 1      Tarvitas dopingut              NA              NA   200
## 2 Ei tarvitanud dopingut              NA              NA  9800
```

Tundlikkus on 25%, seega 25%, kes tarvitas näitab test tarvitajatena. Ehk siis 200*025=50 on tõeste positiivsete testide arv. Ülejäänud (200-50=150) on valenegatiivsed.

```
tabel[1,2]=tabel[1,4]*0.25
tabel[1,3]=tabel[1,4]-tabel[1,2]
tabel
```

```
##                         X Positiivne.test Negatiivne.test KOKKU
## 1      Tarvitas dopingut              50             150   200
## 2 Ei tarvitanud dopingut              NA              NA  9800
```

Spetsiifilisus on 99.9%, seega 99.9% kes, ei tarvitanud dopingut klassifitseerit õigesti õigeteks negatiivseteks.

```
tabel[2,3]=tabel[2,4]*0.999
tabel[2,2]=tabel[2,4]-tabel[2,3]
tabel
```

```
##                         X Positiivne.test Negatiivne.test KOKKU
## 1      Tarvitas dopingut            50.0           150.0   200
## 2 Ei tarvitanud dopingut             9.8          9790.2  9800
```

Tõenäosus, et positiivse testi korral oli sportlane ka dopingu vastu patustaja on õiged positiivsed/(õiged positiivsed+ valed positiivsed). Ehk:

```
tabel[1,2]/(tabel[1,2]+tabel[2,2])
```

```
## [1] 0.8361204
```

- Oletame nüüd, et testi spetsiifilisus on 99.99%. Mis on tõenäosus, et positiivse dopingutesti korral on sportlane ka tegelikult dopingut tarvitanud?

```
tabel[2,3]=tabel[2,4]*0.9999
tabel[2,2]=tabel[2,4]-tabel[2,3]
tabel
```

```
##                         X Positiivne.test Negatiivne.test KOKKU
## 1      Tarvitas dopingut           50.00          150.00   200
## 2 Ei tarvitanud dopingut            0.98         9799.02  9800
```

```
tabel[1,2]/(tabel[1,2]+tabel[2,2])
```

```
## [1] 0.9807768
```

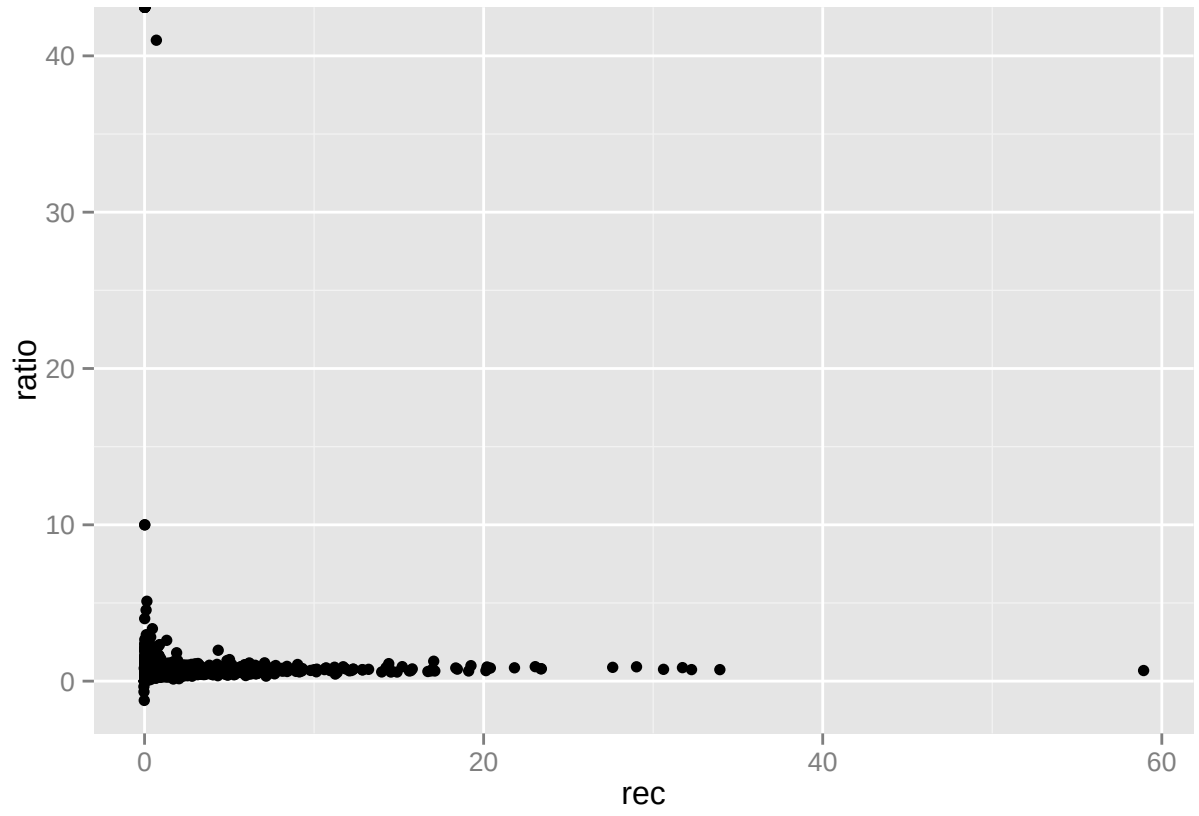
Boonusülesanne (3 punkti)

Kasvuhormooni dopingutesti üks eeldustest oli, et isovormide suhe on konstantne.

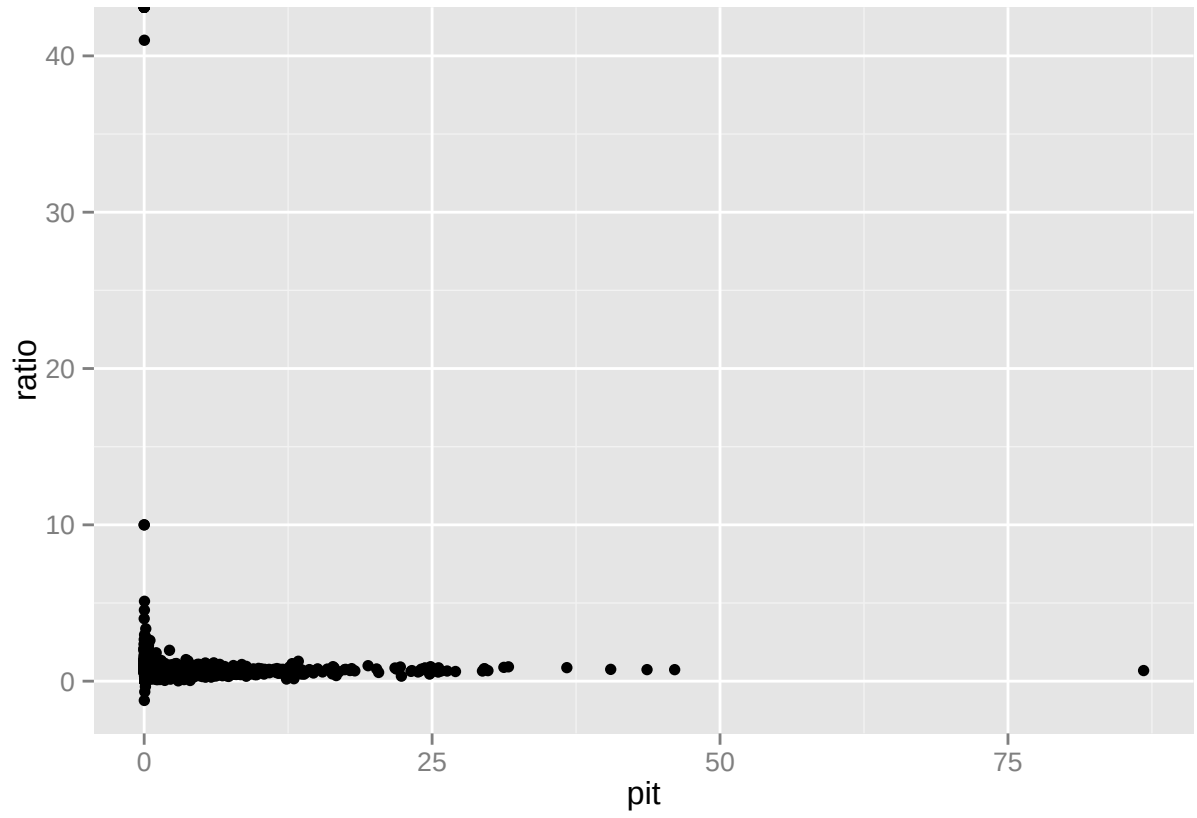
Even though the levels of total hGH concentration will vary substantially, it is assumed the ratio between the relevant types of hGH isoforms measured by the test will naturally remain relatively stable.

Mõtle välja viis, kuidas kontrollida eeldust, et testi mõõdetud isovormide suhe on konstantne ning ei sõltu tegelikust kasvuhormooni kontsentratsioonist. Kontrolli eeldust kasutades enda väljapakutud lähenemist.

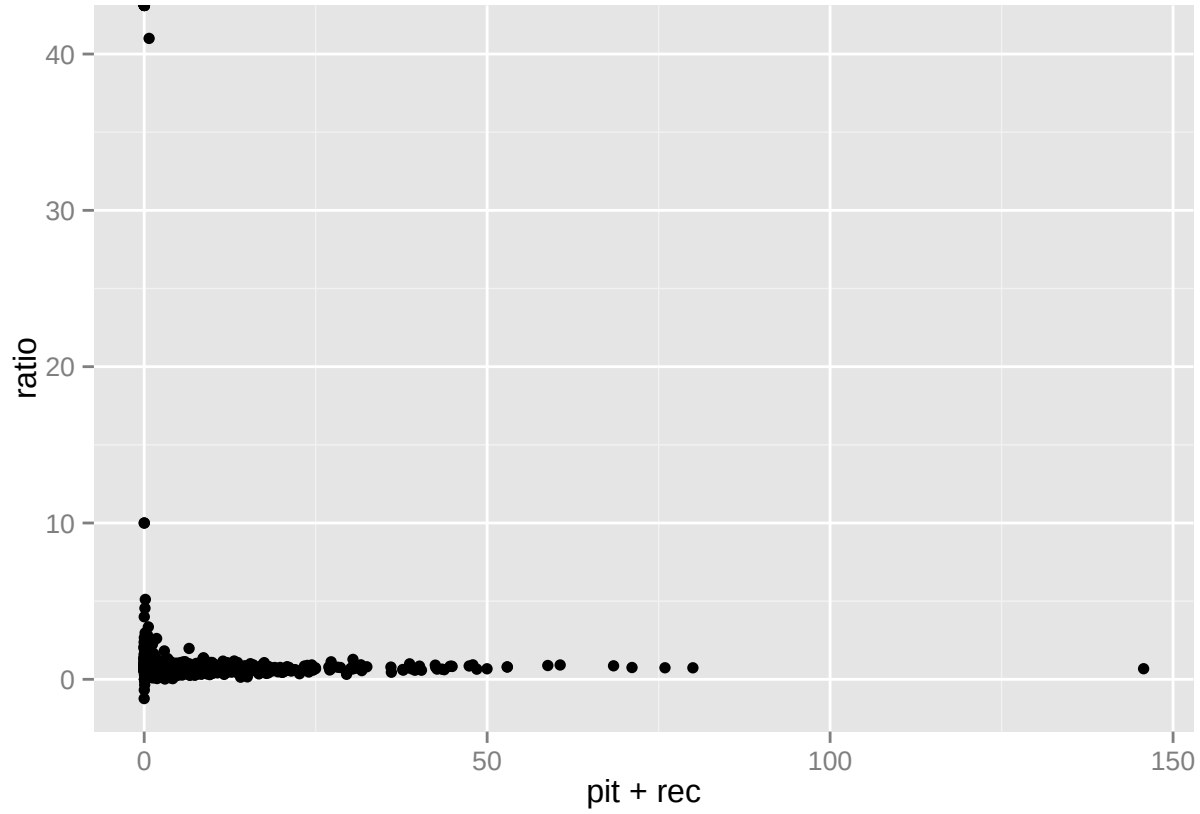
```
library(ggplot2)
#rec vs ratio
ggplot(verification, aes(x=rec, y=ratio))+
    geom_point()
```



```
#pit vs ratio
ggplot(verification, aes(x=pit, y=ratio))+
    geom_point()
```



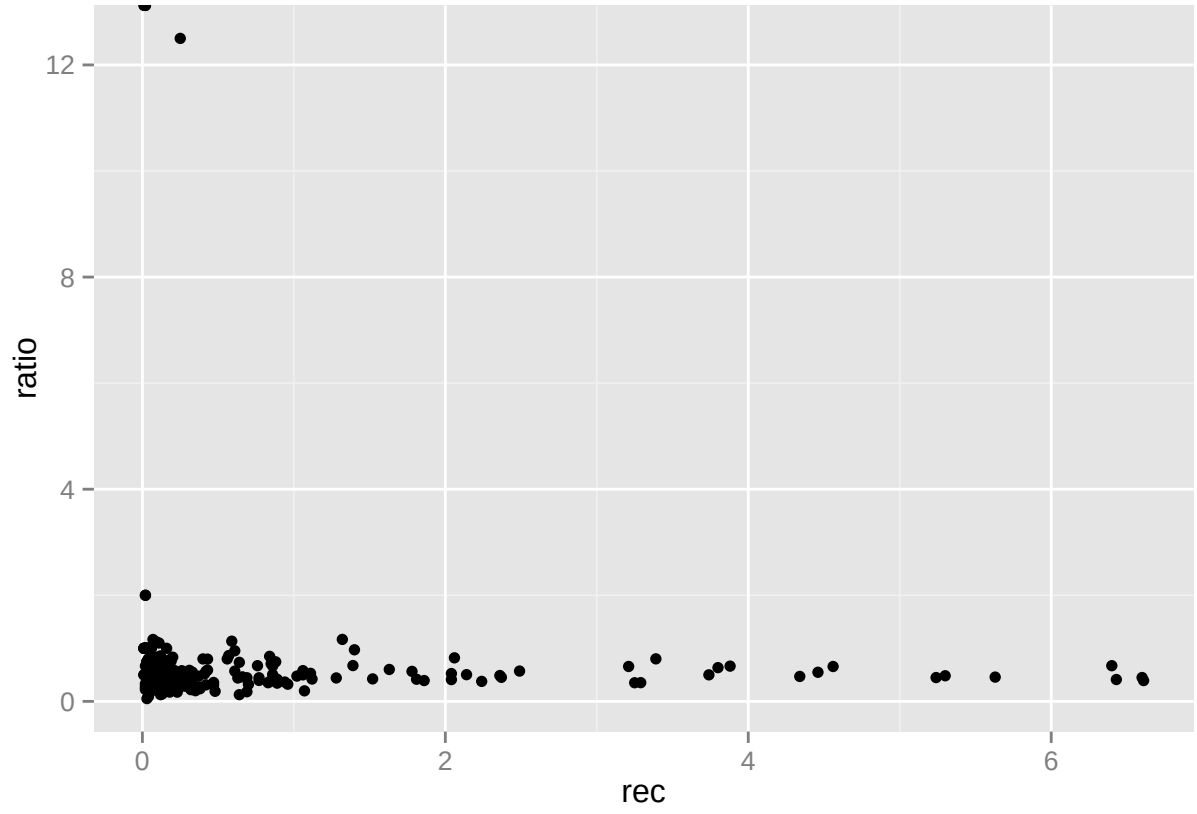

```
#mõlemad liidame
ggplot(verification, aes(x=pit+rec, y=ratio))+
    geom_point()
```



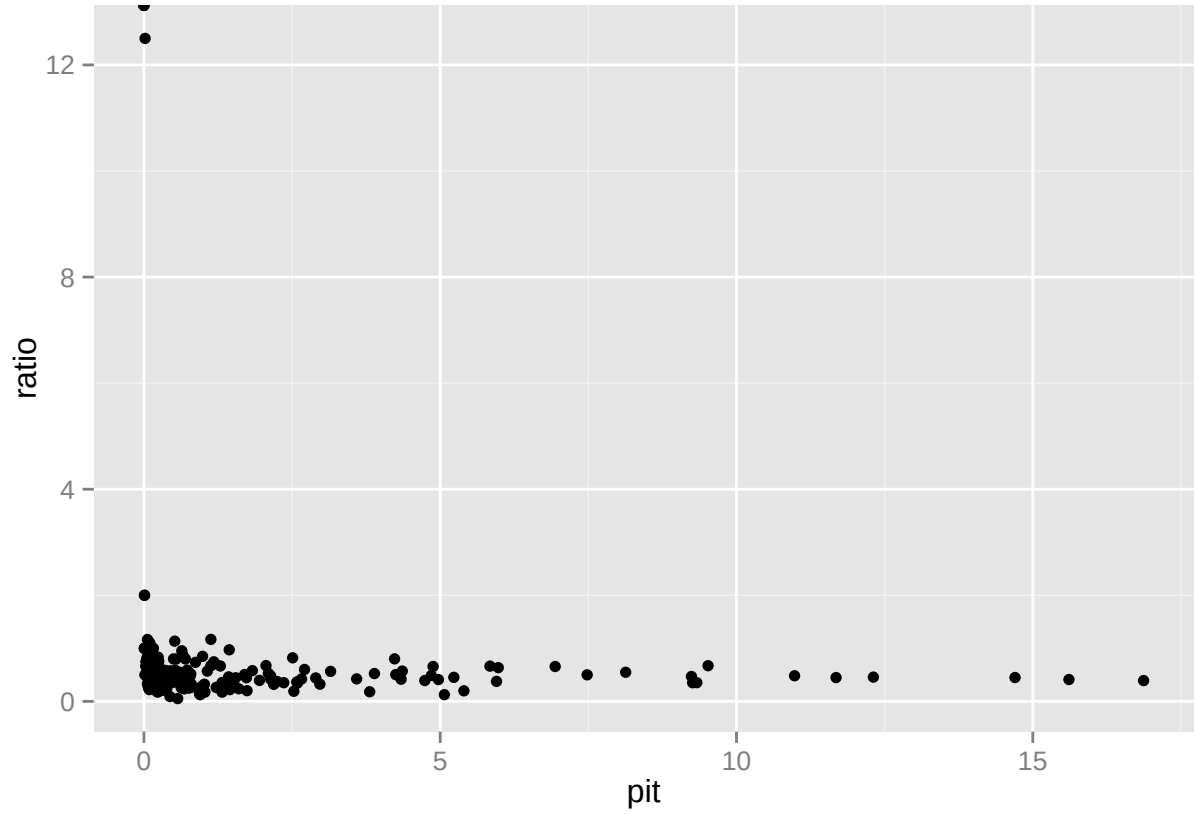

```
#korrelatsioon
verification_puhas=verification[!is.infinite(verification$ratio)&!is.na(verification$ratio),]
cor(verification_puhas$pit, verification_puhas$ratio)
```

```
## [1] 0.01057046
```

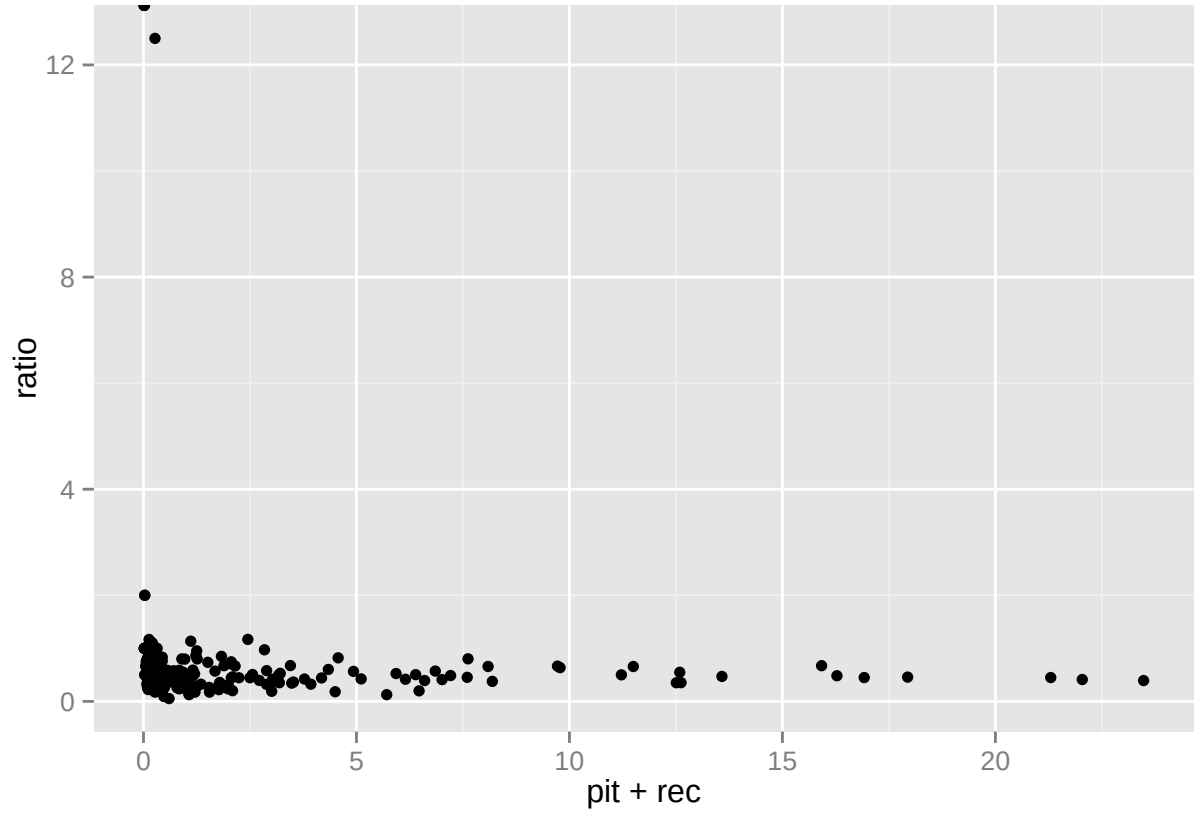
```
#proovin sama algsete andmete pealt
ggplot(doping, aes(x=rec, y=ratio))+
    geom_point()
```




```
#pit vs ratio
ggplot(doping, aes(x=pit, y=ratio))+
    geom_point()
```




```
#mõlemad liidame
ggplot(doping, aes(x=pit+rec, y=ratio))+
    geom_point()
```




```
#korrelatsioon
doping_puhas=doping[!is.infinite(doping$ratio)&!is.na(doping$ratio),]
cor(doping_puhas$pit, doping_puhas$ratio)
```

```
## [1] -0.05677685
```

Proovisin graafilist meetodit ja korrelatsiooni välja arvutada, mõlemad näitavad, et seost pole.