Praktikum 3

R ja selle lisamoodul *Rcmdr*: kirjeldav statistika, joonised, keskmiste võrdlemine

OSA 1
--- Andmestiku avamine jmt *R Commander*'is ---

1) *R Commander*'i käivitamine

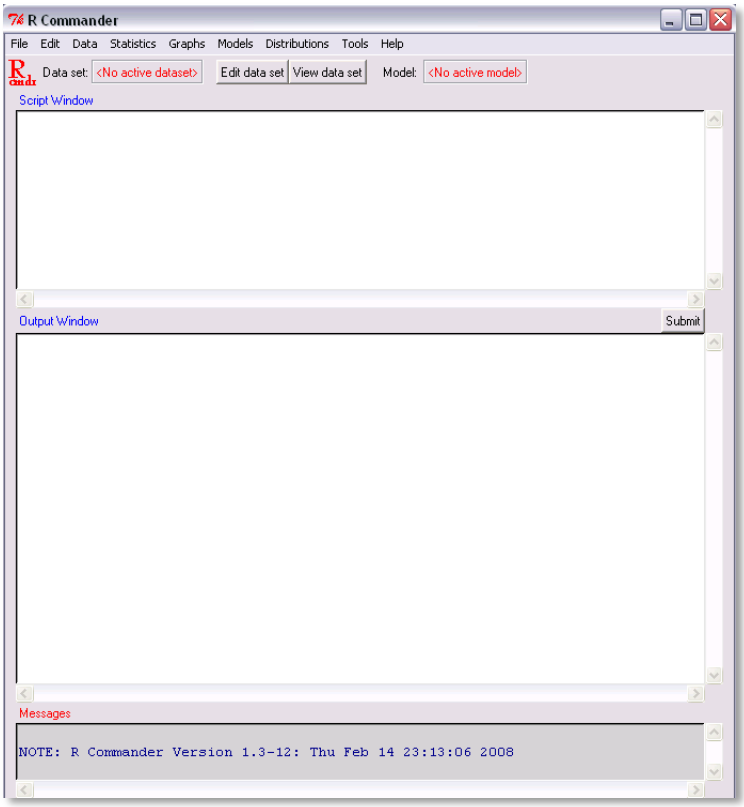
Käivitage *R*.

Käivitage lisamoodul *Rcmdr* trükkides käsureale käsu

```
> library(Rcmdr)
```

või valides vajaliku mooduli *R*-i baas-versioonile lisaks installeeritud pakettide nimekirjast
(*Packages -> Load package ... -> Rcmdr*)

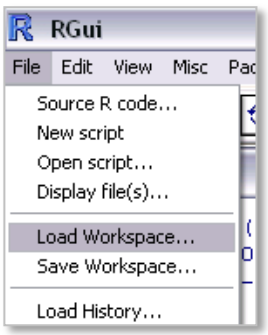
Tulemusena peaks avanema *R Commander*'i aken.



2) Andmete kasutusele võtmine

a) Kui te eelmise praktikumi lõpus salvestasite kõik loodud muutujad (sh *R*-i imporditud andmestiku '*students*') *R workspace*'na (RData-failina), siis avage nimetatud fail:

File -> Load Workspace…



Kui vastavat Rdata-faili pole, vt punkti **2** osasid **b)** ja **c)**.

NB! RData-fail avage *R Commander*'is, mitte *R*'is!

Tulemusena on kasutusvalmis nii kõik eelmisel R-i kasutamisel loodud muutujad kui ka tudengite ankeedivastuste andmestik.

Peale seda kinnitage tudengite andmestik '*students*' *R Commander*'i vaikimisi analüüsitavaks andmestikuks, klikkides nupul

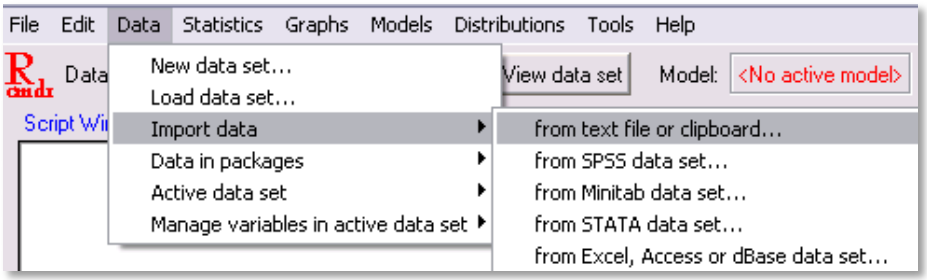
Data set: <No active dataset>

b) Kui teil eelmisest praktikumist salvestatud *R workspace*'i ei ole, siis saate tudengite andmebaasi *R*-i importida ka *R Commander*'i abil
(**seda võivad proovida ka need, kellel nimetatud andmestik eelmisest korrast juba R-s olemas**).

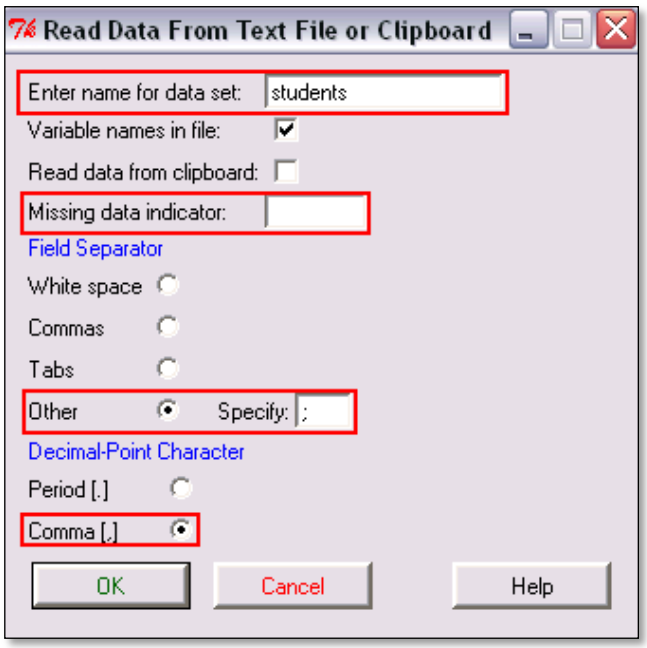
Lihtsaim variant on importida csv-fail otse Internetist aadressilt

http://ph.emu.ee/~ktanel/DK_0007/studentsR.csv

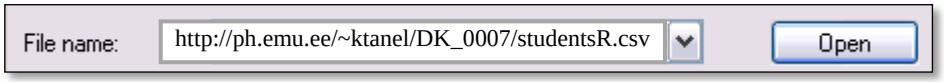
kasutades *R Commander*'i käske



↓



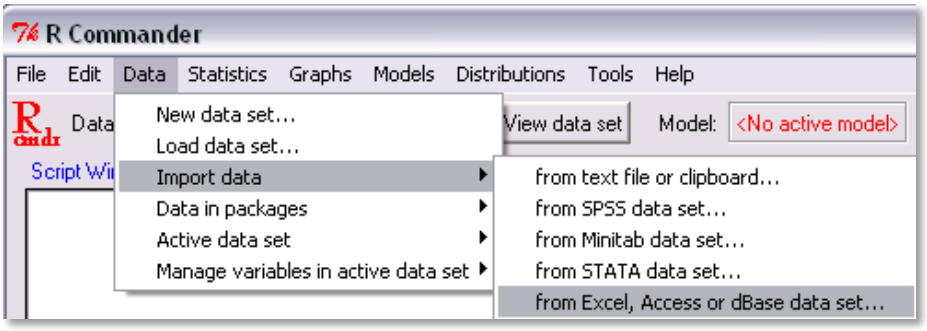
↓ (trüki või kopeeri antud tööjuhendist imporditava faili aadress)



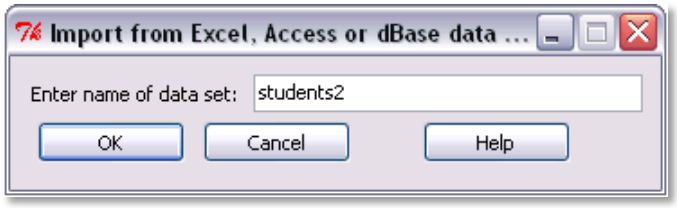
↓ (tulemusena imporditakse andmestik ja muudetakse see automaatselt *R Commander*'is vaikimisi analüüsitavaks)

c) Aga, *R Commander* võimaldab *R*-i importida ka *MS Excel*'i andmetabelit!

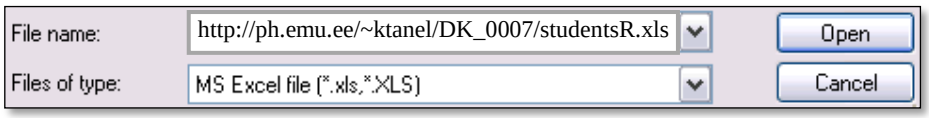
Proovige ka seda varianti
(nimetades uue andmestiku *R*'i jaoks teise nimega, näiteks *students2*).



↓

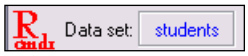


↓ (anna ette kursuse kodulehelt salvestatud *Excel*'i fail või siis Internetiaadress)



Kui imporditav *Excel*'i fail sisaldab mitut töölehte, küsib *R Commander* peale faili asukoha ette andmist täiendavalt, millist töölehte importida.

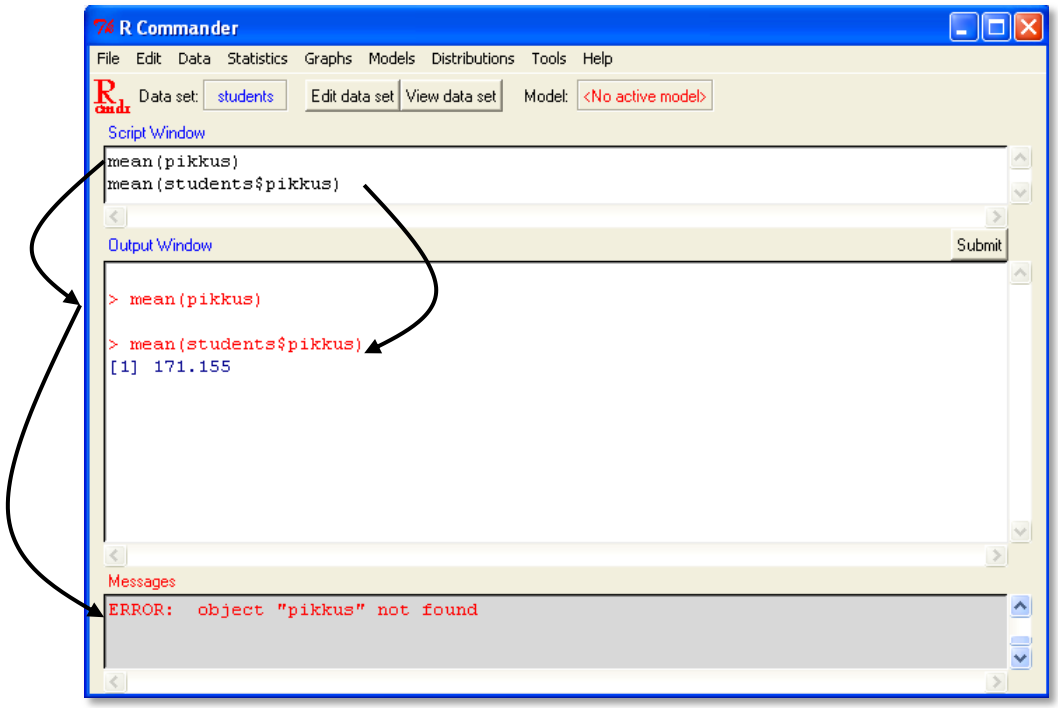
3) Vaikimisi kasutatava andmestiku määratlemine

Valik nupu 

abil kinnitab vastava andmestiku vaikimisi andmestikuks *R Commander*'i menüüde tarvis, aga mitte *R*'i või *R Commander*'i skripti aknast sisestatavate käskude tarvis!

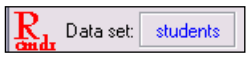
St., et trükkides käsu *R Commander*'i skripti aknasse ja käivitades selle kas klahvikombinatsiooni 'Ctrl'+'R' või nupu 

abil, eeldab *R*, et lisaks analüüsitava tunnuse nimele on ära näidatud ka andmestik, kust antud tunnust otsida.



Kasutamaks tunnuste nimesid skripti aknasse ise kirjutatud käskudes ilma andmestiku nime täpsustamata, tuleb andmestik eelnevalt `attach` käsu abil määrata vaikimisi andmestikuks:

```
attach(students)
```

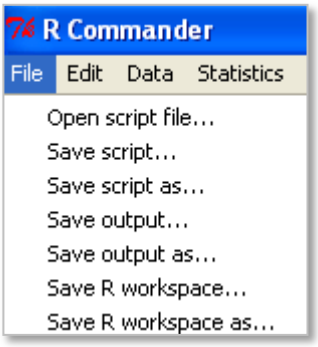
Valik 

R Commander'is **ei ole samaväärne** käsuga `attach(students)`!

4) Salvestamine

a) Skripti salvestamine ja avamine

R Commander'i skripti akna ('*Script Window*') sisu, kuhu *R Commander* genereerib iga menüüdest valitud analüüsi peale ise vastava programmi ning kuhu te saate käske ja kommentaare sarnaselt tavalisele *R*-i skriptiaknale trükkida ka ise, on salvestav *R*'i skriptina
(laiendiga .R, aga oma olemuselt tavalise tekstifailina, mida saate avada ja soovi korral täiendada igas tekstiredaktoris).



Samuti on juba olemasolev skript (näiteks analüüside jätkamiseks) avatav *R Commander*'is.

b) Analüüside tulemuste salvestamine

Tulemuste akna '*Output Window*' sisu, mis koondab kõiki rakendatud käske ja saadud tulemusi, on salvestatav teksti failina (laiendiga .txt).

c) R'i töösessiooni salvestamine

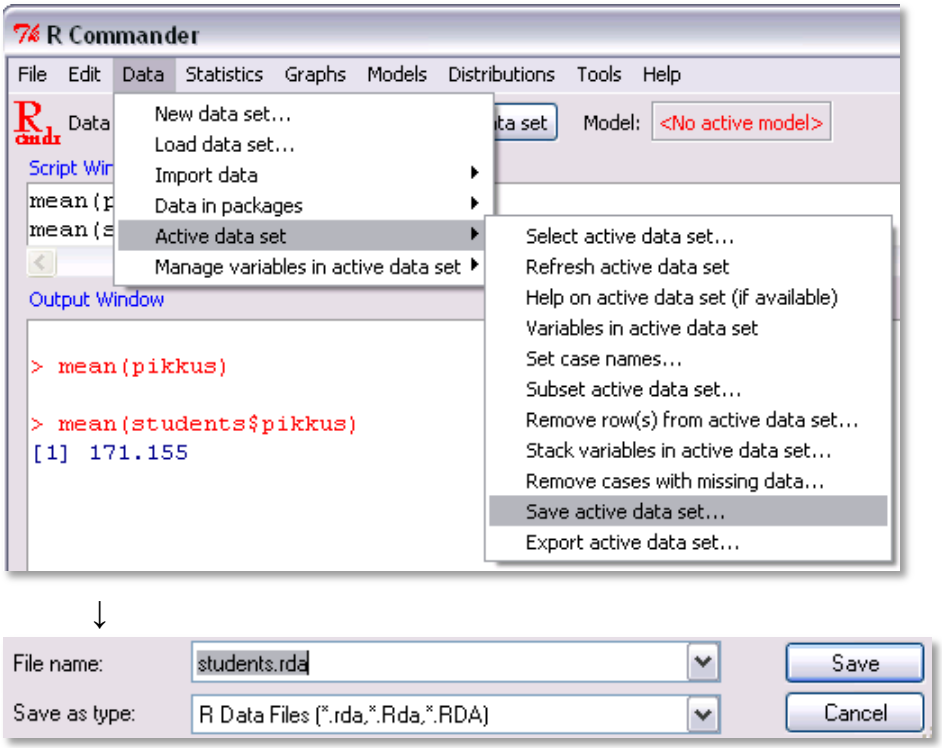
Kõik *R*'i töösessiooni jooksul loodud/imporditud andmestikud ja muutujad on salvestatavad *R workspace*'na (.RData failina).

Järgmisel *R*'i kasutamisel on mugav võtta kõik soovitud andmestikud koheselt kasutusele avades üksnes vastava *R workspace*'i, samuti on vahel arvutuste või modelleerimiste jätkamiseks hea omada korraga hulka juba eelnevalt defineeritud ja väärtustatud muutujaid.

d) Andmete salvestamine

Parajasti aktiivne andmestik (

) on salvestatav ka *R*'i andmetabelina:



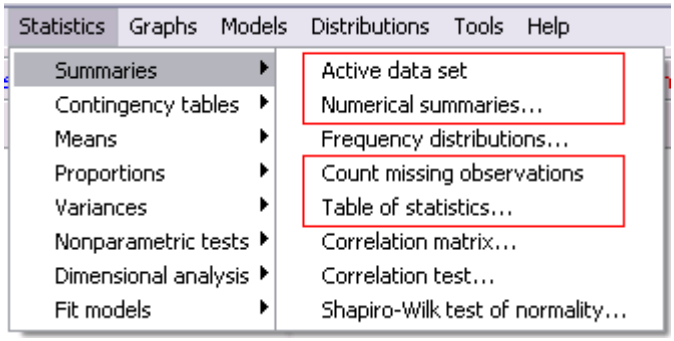
Ja selliselt salvestatud andmestik on jälle *R*'i sisse loetav (*Data -> Load data set...*).

OSA 2
--- Kirjeldav statistika *R Commander*'is ---

1) Rakendage tudengite andmestikule mõningaid *R Commander*'i menüüs

Statistics -> Summaries

sisalduvaid andmetest ülevaate saamiseks ja kirjeldavate statistikute leidmiseks mõeldud (joonisel punase raamiga ümbritsetud) käske.



- Käskude tulemused trükitakse *R Commander*'i väljundiaknasse '*Output Window*'
- ja menüüvalikute tulemusel genereeritud käsud ise *R Commander* skriptiaknasse '*Script Window*'.

Püüdke aru saada, mida need punase raamiga ümbritsetud käsud teevad, millised on nende lisavalikud ja kuidas näevad välja vastavad skriptiaknasse väljastatud käsud.

Püüdke mõningaid skriptiaknasse väljastatud käske muuta (tehes neist eelnevalt soovi korral samasse koopia), rakendamaks neid teistele samas andmestikus sisalduvatele tunnustele või leidmaks teisi arvkarakteristikuid.

***R Commander*'i skriptiaknas sisalduvate käskude rakendamiseks tuleb vajutada 'Ctrl'+'R'**
või siis vajutada nuppu Submit**.**

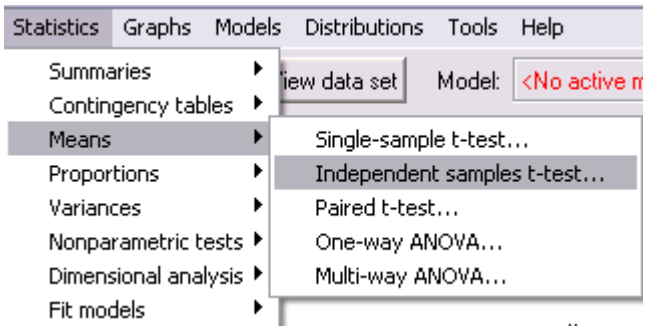
OSA 3
--- Kahe grupi võrdlus *R Commander*'is ---

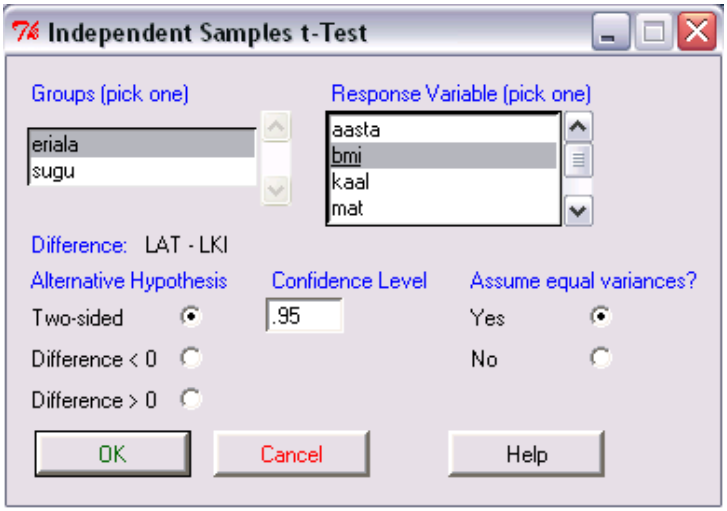
R Commander'i menüüdes sisaldub enamus standardseid kahe grupi võrdlemise teste, lisaks on spetsiifilisemad testid tellitavad lihtsalt vastava käsu skriptiaknasse '*Script Window*' kirjutamisega.

❖ Keskmiste võrdlemine t-testiga

Soovides testida tudengite kehamassiindeksi ('bmi') sõltuvust erialast, tuleks teostada t-test:

Statistics -> Means -> Independent samples t-test …





- Seejuures laseb *R Commander* grupeerivateks tunnusteks ('*Groups*') valida üksnes 2-väärtuselisi mittearvulisi tunnuseid ning uuritavateks tunnusteks ('*Response Variable*'), mille keskmisi võrrelda, andmestikus sisalduvaid arvtunnuseid.
- Lisaks näitab *R Commander*, mis pidi erinevus arvutatakse ('LAT-LKI', sest grupeeriva faktori väärtused esitatab *R* tähestikulises jrk-s),
- võimaldab valida kahepoolset hüpoteesi (t-testi tarvis ise programmi kirjutades funktsiooni `t.test` lisavalik `alternative='two.sided'`) ja kahte erinevat ühepoolset hüpoteesi,
- määrata usaldusnivoo usalduspiiride tarvis (lisavalik `conf.level=`) ning
- teha lisaeelduse dispersioonide võrdumise kohta (lisavalik `var.equal=TRUE`).

```
> t.test(bmi~eriala, alternative='two.sided', conf.level=.95, var.equal=TRUE, data=students)

        Two Sample t-test

data:  bmi by eriala
t = 1.9781, df = 94, p-value = 0.05084
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.00401474  2.15660806
sample estimates:
mean in group LAT mean in group LKI
          22.3613           21.2850
```

Tulemusena arvutab *R* olulisuse tõenäosuse *p* otsustamaks testitud erinevuse statistilise olulisuse üle, leiab keskmiste vahelise erinevuse usaldusintervalli ning väljastab keskmised kehamassiindeksid mõlema eriala tarvis.

Leidsite nimetatud suurused analüüsi väljundist üles? Mida tulemustest järeldate?

R Commander'i skriptiaknasse trükitud käsust ilmneb alternatiivne, universaalsem ja järjest enamate *R*'i funktsioonide poolt aktsepteeritav võimalus võrdlemaks ühes andmetabeli veerus paikneva tunnuse (näiteks 'bmi') väärtusi teises veerus paikneva tunnuse (näiteks 'eriala') väärtuste alusel moodustatud gruppides:

```
bmi ~ eriala
```

Samuti aktsepteerib hulk funktsioone andmestiku ette andmist lisavalikuga:

```
data = students
```

Seega on LAT- ja LKI-eriala tudengite keskmiste kehamassindeksite võrdlemise käsk

```
t.test(students$bmi[students$eriala=='LAT'],students$bmi[students$eriala=='LKI'],
    var.equal=TRUE)
```

alternatiivselt esitatav ka kujul

```
t.test(bmi~eriala, var.equal=TRUE, data=students)
```

(*R Commander*'i poolt funktsiooni `t.test` argumentidena lisatud käske

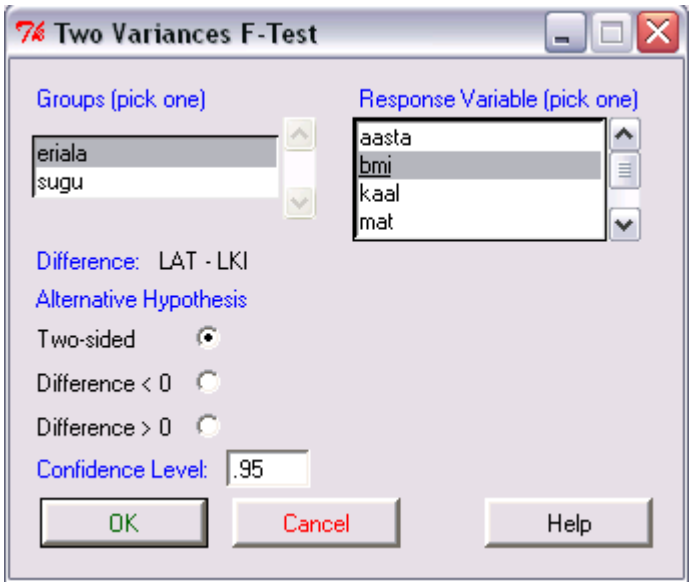
`alternative='two.sided'` ja `conf.level=.95`

pole sellisel kujul tegelikult vaja, sest tegu on vaikimisi väärtustega.)

❖ Dispersioonide võrdlemine

Dispersioonide erinevuse testimiseks kasutatav F-test on leitav menüüst:

Statistics -> Variances -> Two-variances F-test …



↓

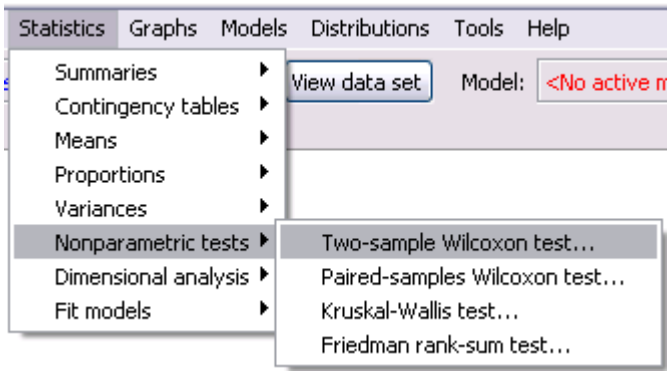
```
> var.test(bmi ~ eriala, alternative='two.sided', conf.level=.95, data=students)

        F test to compare two variances

data:  bmi by eriala
F = 1.3446, num df = 44, denom df = 50, p-value = 0.31
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.7569526 2.4146552
sample estimates:
ratio of variances
          1.344644
```

Tulemuseks olulisuse tõenäosus *p* dispersioonide erinevuse kohta, dispersioonide suhte väärtus ja selle usaldusintervall (kui dispersioonide suhe võrdub 1-ga, ei ole varieeruvus erinev :).

Kas antud juhul oleks põhjust kahelda t-testiga keskmiste võrdlemisel tehtud võrdsete dispersioonide eelduses?

❖ Mitteparameetrilised testid

1) Kahe grupi keskmiste võrdlemisel enam kasutatav mitteparameetriline test – Wilcoxoni test – on *R Commander*'is tellitav menüüst *Statistics -> Nonparametric tests*:



Lisaks tavalisele kahe grupi keskmiste mitteparameetrilisele võrdlemisele on sama menüü samast alajaotusest valitavad testid ka kahe grupi paariviisilise võrdluse tarvis (*Paired-samples Wilcoxon test*) ning enam kui kahe grupi võrdlemise tarvis (dispersioonanalüüsi mitteparameetrilised normaaljaotust mitte eeldavad alternatiivid *Kruskal-Wallis test* ja *Friedman rank-sum test*, neist viimane on paariviisilise mitte-parameetrilise võrdluse üldistus enam kui kahele grupile).

Alternatiivina võib selle testi teostada, trükkides käsureale

```
wilcox.test(students$bmi[students$eriala=='LAT'],students$bmi[students$eriala=='LKI'])
```

või siis

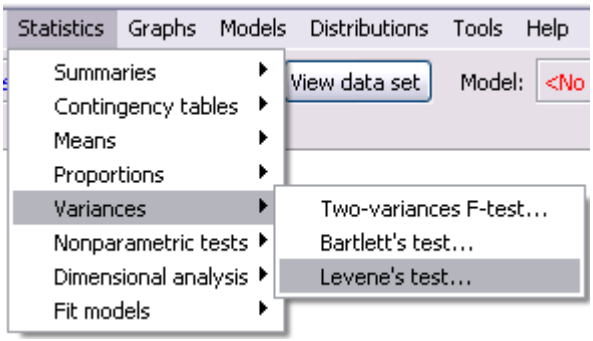
```
wilcox.test(bmi~eriala, alternative="two.sided", data=students)
```

```
> wilcox.test(students$bmi[students$eriala=='LAT'],students$bmi[students$eriala=='LKI'])

        Wilcoxon rank sum test with continuity correction

data:  students$bmi[students$eriala == "LAT"] and students$bmi[students$eriala == "LKI"]
W = 1384.5, p-value = 0.08248
alternative hypothesis: true location shift is not equal to 0
```

Kas LAT- ja LKI-erialade tudengite kehamassiindekseid on põhjust erinevaiks lugeda?

---

2) Nii kahe kui ka enama grupi dispersioonide mitte-parameetriliseks normaaljaotust mitte eeldavaks võrdlemiseks sisaldub *R Commander*'is Levene' test (see test on samuti enam kui kahe grupi dispersioonide võrdlemiseks, aga normaal-jaotuse eeldustel, mõeldud Bartlett' testi üldistus).



Teostage eelnevalt F-testiga sooritatud LAT- ja LKI-eriala tudengite kehamassiindeksite dispersioonide võrdlus ka Levene' testiga.

On's lõppjäreldus sama – dispersioonid ei ole erinevad?

```
> tapply(students$bmi, students$eriala, var, na.rm=TRUE)
     LAT      LKI
8.194334 6.094057

> levene.test(students$bmi, students$eriala)
Levene's Test for Homogeneity of Variance
      Df F value Pr(>F)
group  1  0.0015 0.9696
      94
```

NB! Pane tähele funktsiooni

`levene.test`

süntaksi (argumentide esituse) erinevust teistest siiani vaadeldud testidest!

Veel üks normaaljaotust mitte eeldav dispersioonide võrdlemise test on Fligner-Killeen'i test. See test ei ole tellitav *R Commander*'i menüüdest (nagu tegelikult suur hulk teisigi teste – *R*'is olla üle 100 kahe grupi võrdlemiseks kasutatava testi …), aga on rakendatav, trükkides vastava käsu lihtsalt skriptiaknasse ja käivitades sealt (seejuures on funktsiooni `fligner.test` süntaks analoogne funktsiooni `levene.test` omaga):

```
fligner.test(students$bmi, students$eriala)
```

```
> fligner.test(students$bmi, students$eriala)

        Fligner-Killeen test of homogeneity of variances

data:  students$bmi and students$eriala
Fligner-Killeen:med chi-squared = 0.0474, df = 1, p-value = 0.8277
```

NB! Tegelikult mõistab *R* nii Levene' kui ka Fligner-Killeen'i testi teha ka käskude

```
fligner.test(bmi ~ eriala, data=students)
```

ja

```
levene.test(bmi ~ eriala, data=students)
```

alusel. Ainult et mingil põhjusel *R Commander* Levene' testi puhul taolist süntaksit erinevalt teistest testidest ise ei kasuta …

3) Uuritava tunnuse jaotuse võrdlemiseks kahes grupis (normaaljaotust eeldamata) on kasutatav Kolmogorov-Smirnovi test. Ka see test ei ole realiseeritud *R Commander*'i menüüdes, mistõttu tuleb käivitada skriptiaknast:

```
ks.test(students$bmi[students$eriala=='LAT'],
                              students$bmi[students$eriala=='LKI'])
```

```
> ks.test(students$bmi[students$eriala=='LAT'],students$bmi[students$eriala=='LKI'])

        Two-sample Kolmogorov-Smirnov test

data:  students$bmi[students$eriala == "LAT"] and students$bmi[students$eriala == "LKI"]
D = 0.2758, p-value = 0.05264
alternative hypothesis: two-sided
```

Et p = 0,053, siis ei ole LAT- ja LKI-eriala tudengite kehamassiindeksite jaotuse erinevus statistiliselt oluline.

❖ Jaotuste võrdlemine

1) Lisaks kahe tunnuse jaotuste omavahelisele võrdlemisele on Kolmogorov-Smirnovi test kasutatav ka uuritava tunnuse jaotuse võrdlemiseks normaaljaotusega (või mõne teise teoreetilise jaotusega):

```
ks.test(students$bmi, pnorm, mean=mean(students$bmi,na.rm=TRUE),
                                    sd=sd(students$bmi,na.rm=TRUE))
```

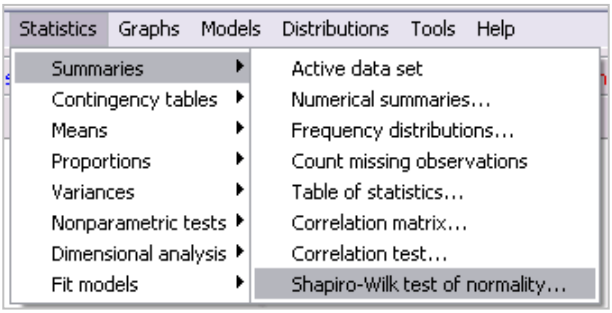
```
> ks.test(students$bmi, pnorm, mean=mean(students$bmi,na.rm=TRUE), sd=sd(students$bmi,na.rm=TRUE))

        One-sample Kolmogorov-Smirnov test

data:  students$bmi
D = 0.0857, p-value = 0.4804
alternative hypothesis: two-sided
```

Järeldus: kuna $p = 0{,}48 > 0{,}05$, siis pole Kolmogorov-Smirnovi testi alusel põhjust ümber lükata nullhüpoteesi tudengite kehamassiindeksite normaaljaotuse järgsest jaotumisest.

2) Alternatiivina võib tunnuse normaaljaotuse järgset jaotumist testida ka Shapiro-Wilksi testiga:

```
shapiro.test(students$bmi)
```

See test on tellitav ka *R Commander*'i menüüst:



Tulemus:

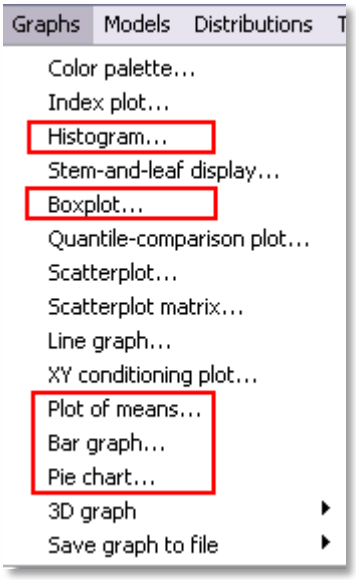
```
> shapiro.test(students$bmi)

        Shapiro-Wilk normality test

data:  students$bmi
W = 0.962, p-value = 0.00699
```

Nagu ka tulemustest nähtub, on Shapiro-Wilksi test märgatavalt tundlikum uuritava tunnuse jaotuse kõrvalekallete suhtes normaalsusest – näiteks antud juhul tuleks erinevalt Kolmogorov-Smirnovi testist lugeda tõestatuks tudengite kehamassiindeksite jaotuse erinevus normaaljaotusest ($p = 0{,}007$).

Erinevate testide erinevate tulemuste põhjus on erinevas arvutamismetoodikas ja sellest lähtuvalt testide erinevas ranguses. Shapiro-Wilksi test on spetsiaalselt konstrueeritud võrdlemaks uuritava tunnuse jaotust normaaljaotusega ja testib seega mitmeid üksnes normaaljaotusega seonduvaid aspekte, Kolmogorov-Smirnovi test on aga üliuniversaalne ja seetõttu ka üsna robustne jaotuste võrdlemisel kasutatav test, mis võimaldab normaaljaotusega võrdlemist lihtsalt ühena paljudest valikutest ega ole seetõttu suuteline tungima normaaljaotuse peensustesse.

OSA 4
--- Joonised *R Commander*'is ---

Püüdke *R Commander*'i abil konstrueerida alloleval menüüpildil punasega ümbritsetud jooniseid.



NB! *R Commander*'i menüüdest tellitud jooniste tarvis kirjutatakse vastav programm *R Commander*'i skripti aknasse (kus te saate seda muuta/täiendada ja uuesti juba ilma menüüdest valimata käivitada.

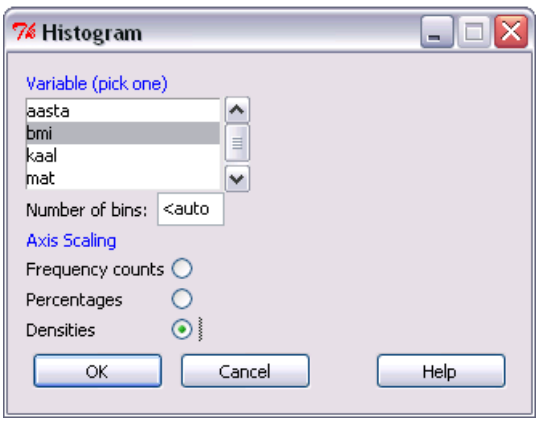
Joonised ise tehakse *R*'is (mitte *R Commander*'is!) avanevasse jooniste aknasse.

Ütlemaks *R*'ile, et ta vanu jooniseid uutega üle ei kirjutaks, tuleb *R*'i aknas menüüst *History* valida käsk *Recordings*:

Püüdke *R Commander*'i poolt vaikimisi produtseeritavaid jooniseid ilusamaks teha, täiendades skriptiaknas jooniste tegemise käske eelmises praktikumis vaadeldud lisavalikutega (lisamaterjali jooniste parameetrite tarvis vt ka http://www.ms.ut.ee/mart/R/Rgraafika.html).

Konstrueerige näiteks tudengite kehamassiindeksite histogramm, valides y-telje skaalaks jaotustiheduse:

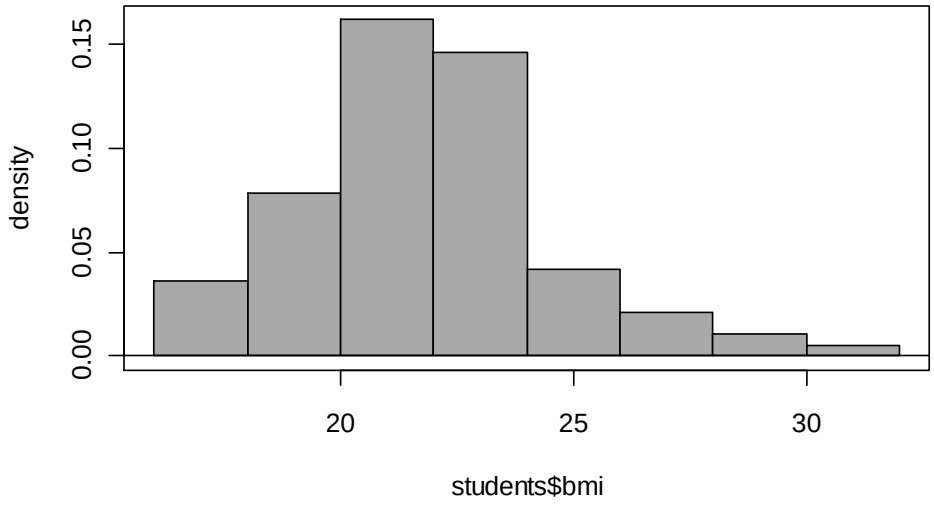
Graphs -> Histogram ->



Tulemusena tekitab *R Commander* skriptiaknasse käsu

```
Hist(students$bmi, scale="density", breaks="Sturges", col="darkgray")
```

ja *R*'i aknasse järgmisel leheküljel toodud joonise.



R Commander'i funktsioon `Hist` on eelmises praktikumis rakendatud *R*'i baasversiooni kuuluva funktsiooni `hist` pisut täiendatud versioon[1].

Sellele saab skriptiaknas lisada erinevaid argumente, parandamaks tulemuseks oleva joonise informatiivsust ja välja nägemist.

Näiteks modifitseerides *R Commander*'i poolt skriptiaknasse kirjutatud käsku kujul

```
Hist(students$bmi,scale="density",xlab="Kehamassiindeks",ylab="Jaotustihedus",xlim=c(14,32))
lines(density(na.omit(students$bmi)), lwd=2)

x=seq(12, 32, length=300)
y=dnorm(x, mean=mean(students$bmi,na.rm=T), sd=sd(students$bmi,na.rm=T))
lines(x, y, lwd=2, col="red")
```

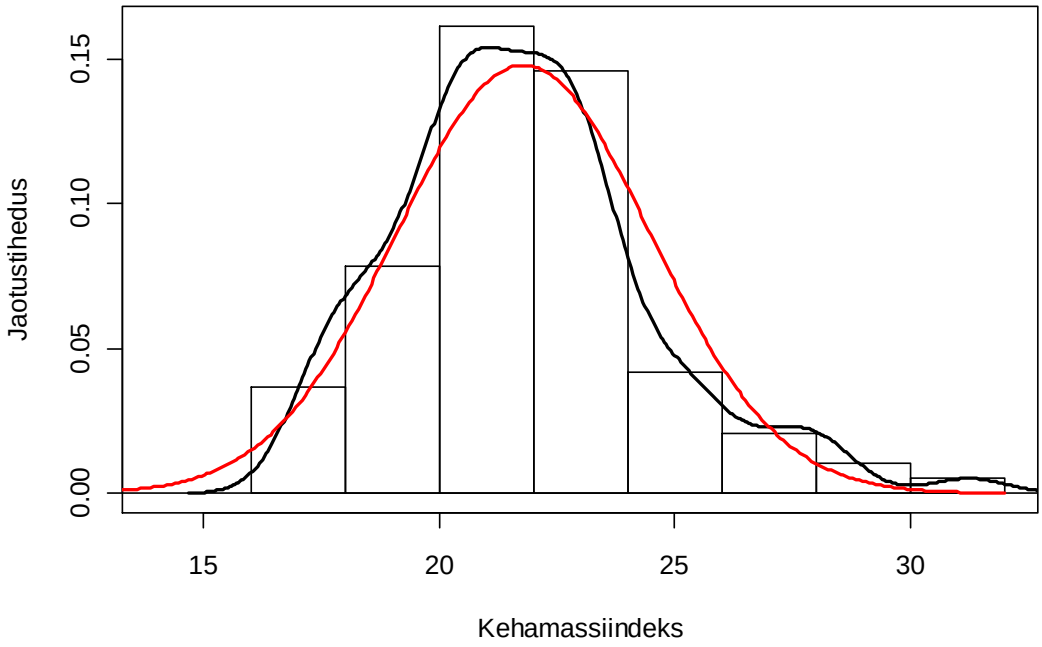
saame tulemuseks joonise, kus algsele histogrammile on lisatud tudengite kehamassiindeksi empiirilise tihedusfunktsiooni graafik (musta joonena) ja vastava (sama keskmise ja standardhälbega) normaaljaotuse tihedusfunktsiooni graafik (punase joonena), samuti on lisatud korrektsed telgede nimed ja muudetud x-telje ulatust.

[1] Paljud *R Commander*'i funktsioonid on ehitatud *R*'i baasversiooni vastavate funktsioonide peale, sageli on neil nimedki samad, *R Commander*'i funktsiooni eristab selle aluseks olevast funktsioonist üksnes suur esitaht funktsiooni nimes (näiteks `Hist` vs `hist`).

Kõik *R*'i baasversioonile omased lisavalikud ja argumendid töötavad ka selle peale ehitatud *R Commander*'i funktsiooni korral, samas võib *R Commander*'i funktsioon sisaldada üksnes sellega töötavaid lisavalikuid.

Näiteks funktsiooni `hist` lisavalikud `freq=TRUE` ja `freq=FALSE` töötavad korrektselt ka funktsiooni `Hist` korral, aga *R Commander*'i funktsiooni `Hist` sama tulemuse andvad alternatiivsed lisakäsud `scale="frequency"` ja `scale="density"` *R*'i baasversiooni kuuluva funktsiooniga `hist` ei tööta.

Lisaks võib *R Commander*'i alternatiivsetel lisakäskudel olla ka täiendavaid väärtusi, näiteks on võimaldab funktsiooni `Hist` lisakäsk `scale="percent"` esitada histogrammi y-telje ühikud protsentides.


Jaotustihedus
0.00
0.05
0.10
0.15
15
20
25
30
Kehamassiindeks