R-tölfræðiúrvinnsla

Seinni hluti

Anna Helga Jónsdóttir
Sigrún Helga Lund

Helstu atriði:

1. Samfelldar líkindadreifingar
2. Ályktanir um flokkabreytur
3. Ályktanir um talnabreytur
4. Aðhvafsgreining
5. Fervikagreining (ANOVA)
6. Ákvörðunartré
7. Tímaraðir
8. Íslenskir stafir

Notum púlsgögnin eins og áður

Byrjum á því að lesa inn púlsgögnin og kóða kyn og líkamsræktarbreytuna:

```
puls<-read.table("pulsAll.csv",header=T,sep=";")
```

```
puls$kyn <- factor(puls$kyn,levels=c(1,2),labels = c("Kona","Karl"))
puls$likamsraektf <- cut(puls$likamsraekt, c(0, 2, 5, 26),right=F)
levels(puls$likamsraektf) <- c('lítil','miðlungs','mikil')
```

Við munum svo nota ggplot2 og tidyr pakkana. Gott að ná í þá strax:

```
library(ggplot2)
library(dplyr)
```

Hvert erum við komin...

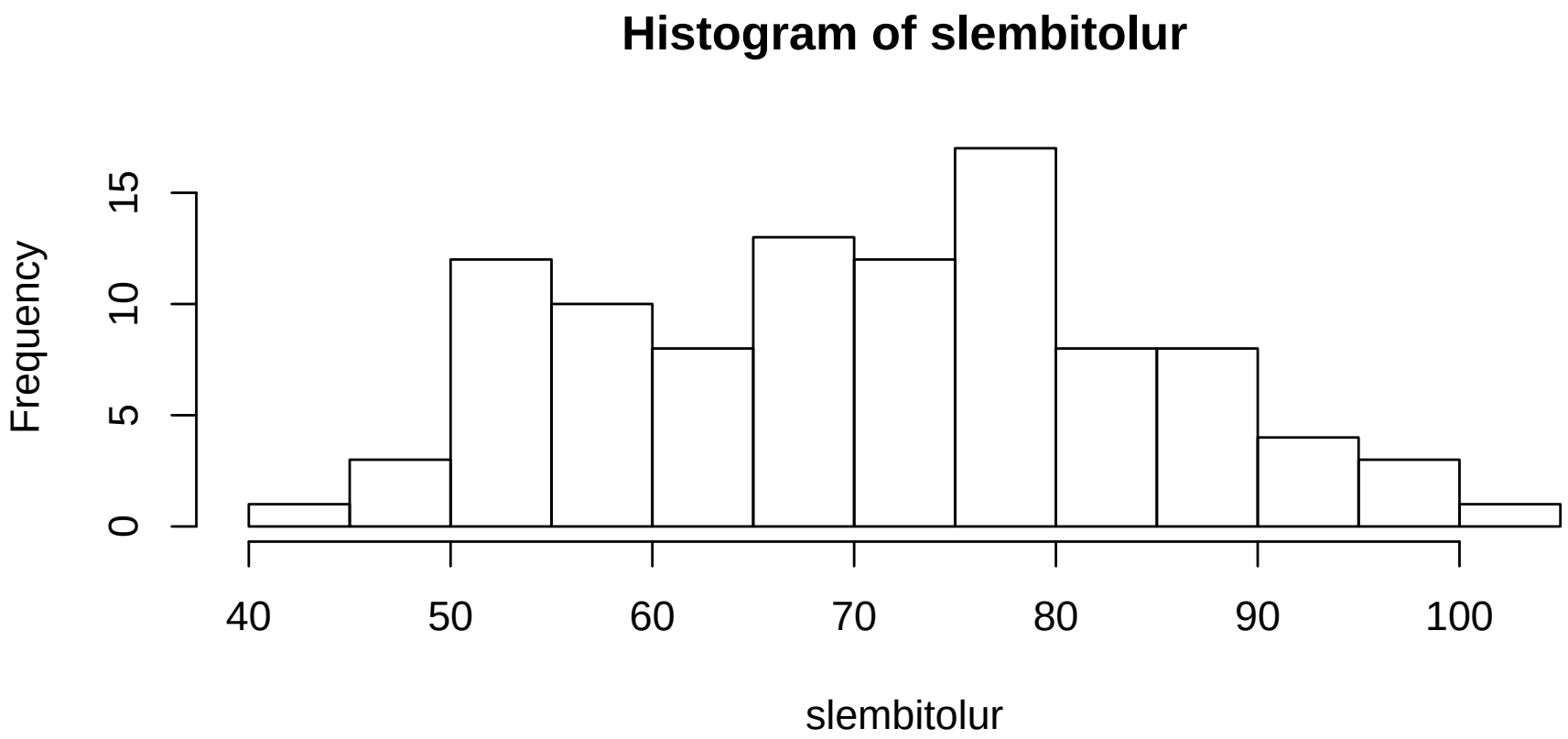
Samfelldar líkindadreifingar

- Fyrir allar helstu samfelldar líkindadreifingar eru til þrjár gerðir aðferða:
  - `p`- skilar okkur *dreififalli* (*probability distribution funciton*)
  - `q`- skilar okkur *hlutfallsmörkum* (*quantiles*)
  - `r`- skilar okkur slembitölu úr dreifingunni.
- Notum normaldreifinguna sem dæmi hér en sambærilegar aðferðir til fyrir t-, F-, kí-kvaðrat dreifinguna, ...

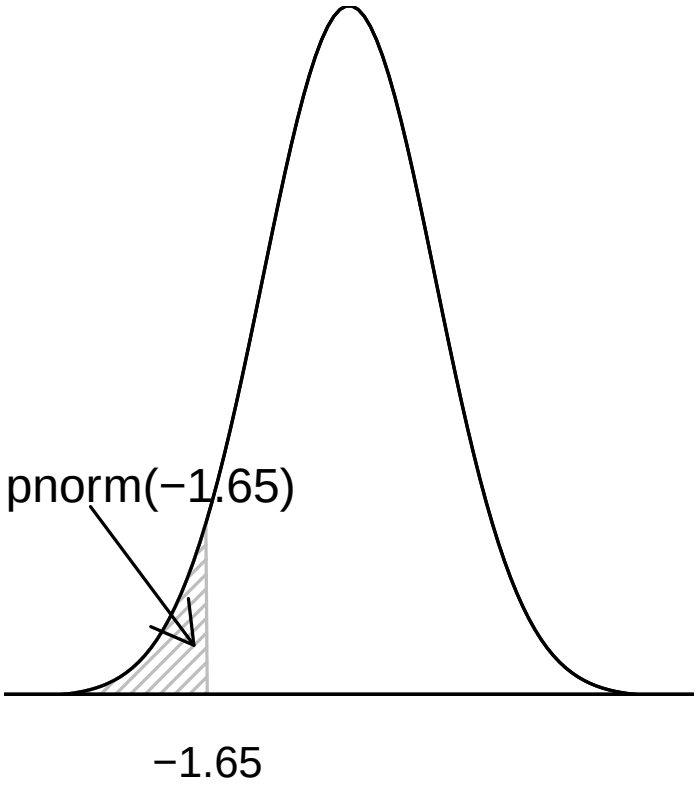
rnorm()

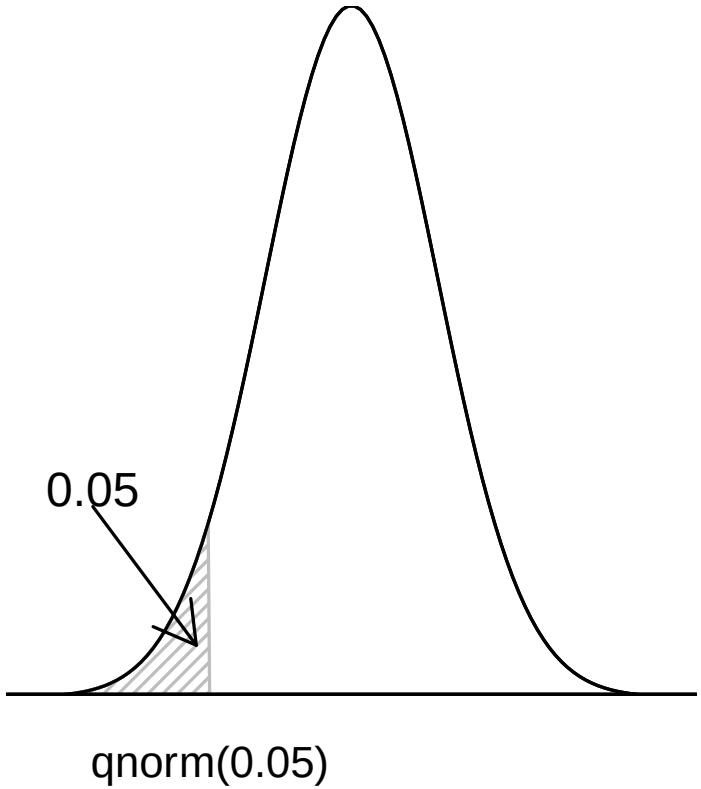
Skipunin `rnorm()` býr til tölur sem fylgja normaldreifingu.

```
slembitolur <- rnorm(100, mean=72, sd=12 )
hist(slembitolur)
```



Munur á p- og q-





pnorm()

Skipunin `pnorm()` reiknar líkurnar á því að normaldreifð slembistærð taki gildi sem er minna en viðmiðunargildi sem við mötum fallið á.

Líkurnar á því að normaldreifð slembistærð með meðaltal 72 og staðalfrávik 12 taki gildi sem er minna en viðmiðunargildið 51 er

```
pnorm(51, mean=72,sd=12)

## [1] 0.04005916
```

eða u.þ.b. 4%.

qnorm()

Skipunin `qnorm()` finnur viðmiðunargildi fyrir gefnar líkur.

Það viðmiðunargildi sem er þannig að líkurnar á því að fá það gildi, eða minni tölu, eru jafnar gefnu líkunum.

Það viðmiðunargildi sem er þannig að það eru 4% líkur á að fá gildi sem er í mesta lagi svo stórt út úr slembistærð með meðaltal 72 og staðalfrávik 12 er

```
qnorm(0.04, mean=72,sd=12)

## [1] 50.99177
```

eða u.þ.b. 51.

Hvert erum við komin...

Ályktanir um hlutfall í einu þýði

- Við notum aðferðina `binom.test()` til að kanna tilgátur og smíða öryggisbil fyrir hlutfall þýðis.
- Skipunin er mötuð á fjölda útkoma af hvorri gerð sem fá má með `table()` skipuninni.
- Notum aðferðina til að kanna hvort kynjahlutfallið sé jafnt í nemendahópnum sem púlsgögnin byggja á.

Ályktanir um hlutfall í einu þýði

Byrjum á því að nota `prop.table()` skipunina, til að sjá hvert kynjahlutfallið er:

```
prop.table(table(puls$kyn))

##
##      Kona      Karl
## 0.6518047 0.3481953
```

Tilgátuprófið, öryggisbilið, prófstærðin og tilgátuprófið fást öll í einu með einni skipun:

Ályktanir um hlutfall í einu þýði

```
binom.test(table(puls$kyn))

##
##  Exact binomial test
##
## data:  table(puls$kyn)
## number of successes = 307, number of trials = 471, p-value =
## 4.308e-11
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.6068703 0.6948191
## sample estimates:
## probability of success
##              0.6518047
```

Ályktanir um hlutföll í tveimur eða fleiri þýðum

- Viljum við draga ályktanir um tvö eða fleiri hlutföll getum við ekki lengur notað `binom.test()` skipunina.
- Þess í stað notum við skipunina `prop.test()` sem byggir á normalnálgun.
- Hún gefur sömu niðurstöðu og algengar aðferðir sem hægt er að reikna í höndunum og eru kenndar í flestum kennslubókum.

Ályktanir um hlutföll í tveimur eða fleiri þýðum

Byrjum á því að skoða hvert kynjahlutfallið er með `prop.table()`.

```
prop.table(table(puls$kyn, puls$namskeid),margin=2)

##
##            LAN203   STAE209
##   Kona 0.6453488 0.6555184
##   Karl 0.3546512 0.3444816
```

Ályktanir um hlutföll í tveimur eða fleiri þýðum

Tilgátuprófið, öryggisbilið, prófstærðin og tilgátuprófið fást öll í einu með einni skipun:

```
prop.test(table(puls$namskeid, puls$kyn))

##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  table(puls$namskeid, puls$kyn)
## X-squared = 0.015035, df = 1, p-value = 0.9024
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.10426340  0.08392429
## sample estimates:
##    prop 1    prop 2
## 0.6453488 0.6555184
```

Ályktanir um hlutföll í tveimur eða fleiri þýðum

Einnig er hægt að nota skipunina `prop.test()` til að bera saman hlutföll fleiri en tveggja hópa. Þá þarf að gæta þess að tengslataflan snúi rétt:

```
prop.test(table(puls$likamsraektf, puls$kyn))

##
##  3-sample test for equality of proportions without continuity
##  correction
##
## data:  table(puls$likamsraektf, puls$kyn)
## X-squared = 11.3, df = 2, p-value = 0.003518
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3
## 0.6117647 0.7382199 0.5789474
```

Ályktanir um tengslatöflur

- Viljum við kanna hvort samband sé á milli tveggja flokkabreyta er notuð aðferðin `chisq.test()` aðferðinni.
- Þá skipun er einni hægt að nota til að bera saman hlutföll tveggja eða fleiri þýða en hún gefur að vísu ekki öryggisbil eins og `prop.test()` skipunin.

Ályktanir um tengslatöflur

Könnum nú hvort samband sé á milli námskeiðs og líkamsræktarástundunar. Við mötum aðferðina með mældri tíðni sem fæst með `table()` skipuninni:

```
chisq.test(table(puls$namskeid,puls$likamsraektf))

##
##  Pearson's Chi-squared test
##
## data:  table(puls$namskeid, puls$likamsraektf)
## X-squared = 4.1576, df = 2, p-value = 0.1251
```

Ályktanir um tengslatöflur

Munið að til þess að geta notað kí-kvaðrat prófið þurfa allar tölurnar í væntitíðnitöflunni að vera stærri en 5. Við getum fengið væntitíðnitöflu út úr R með að vista það sem `chisq.test()` aðferðin skilar sem hlut (hann má heita hvað sem er) og draga svo `expected` hlutann fram:

```
kikv1<-chisq.test(table(puls$namskeid,puls$likamsraektf))
kikv1$expected

##
##              lítil  miðlungs     mikil
##   LAN203  31.19099  70.08798  69.72103
##   STAE209 53.80901 120.91202 120.27897
```

Hvert erum við komin...

Dreifni í tveimur þýðum

Við mötum `var.test()` aðferðina á mismunandi hátt eftir því hvort gögnin okkar séu á löngu eða víðu sniði.

Viljum við kanna hvort dreifni í púls sé mismunandi á milli kynjanna gerum við það með (takið eftir að gögnin eru á löngu sniði):

```
var.test(puls$fyrriPuls~puls$kyn)


##
##  F test to compare two variances
##
## data:  puls$fyrriPuls by puls$kyn
## F = 1.2512, num df = 293, denom df = 159, p-value = 0.1155
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9461411 1.6360633
## sample estimates:
## ratio of variances
##           1.251186
```

Dreifni í fleiri þýðum

Þegar þýðin/hóparnir eru fleiri en tveir má nota `bartlett.test()` aðferðina. Við þurfum að mata aðferðina með vigri sem inniheldur mælingarnar okkar og annan vigur sem tilgreinir hvaða hópi mælingarnar tilheyra.

```
bartlett.test(puls$fyrriPuls ~ puls$likamsraektf)

##
##  Bartlett test of homogeneity of variances
##
## data:  puls$fyrriPuls by puls$likamsraektf
## Bartlett's K-squared = 4.1246, df = 2, p-value = 0.1272
```

Ályktanir um meðaltal í einu þýði

Þegar skipunin `t.test()` er mötuð með einungis einni breytu framkvæmir hún t-próf fyrir eitt meðaltal. Aðrar stillingar eru:

- `mu`: Við prófum tilgátuprófið $H_0 : \mu =$ `mu`. `mu` er því viðmiðunargildi núlltilgátunnar.
- `alternative`: Við gefum skipunina `alternative=''two.sided''` ef gagntilgátan er tvíhliða, `alternative=''greater''` ef gagntilgátan er $\mu > \mu_0$ og `alternative=''less''` ef gagntilgátan er $\mu < \mu_0$. Sjálfgefið er að hafa tvíhliða gagntilgátu.
- `conf.level`: Þar tilgreinum við hvert öryggið (og þá um leið villulíkurnar) á að vera fyrir tilgátuprófið og öryggisbilið. Sjálfgefið er að hafa öryggið $1 - \alpha = 0.95$.

Ályktanir um meðaltal í einu þýði

Könnum hvort púlsinn sé frábrugðinn 70:

```
t.test(puls$fyrriPuls,mu=70)

##
##  One Sample t-test
##
## data:  puls$fyrriPuls
## t = 3.5612, df = 453, p-value = 0.0004082
## alternative hypothesis: true mean is not equal to 70
## 95 percent confidence interval:
##  70.88843 73.07633
## sample estimates:
## mean of x
##  71.98238
```

Ályktanir um mismun meðaltala tveggja óháðra þýða

Við mötum `t.test()` aðferðina á mismunandi vegu eftir því á hvaða sniði gögnin eru. Ennfremur er hægt að gefa eftirfarandi stillingar

- `mu`: Við prófum tilgátuprófið $H_0 : \mu_1 - \mu_2 =$ `mu`. `mu` er því viðmiðunargildi núlltilgátunnar.
- `conf.level`: Þar tilgreinum við hvert öryggið (og þá um leið villulíkurnar) á að vera fyrir tilgátuprófið og öryggisbilið. Sjálfgefið er að hafa öryggið $1 - \alpha = 0.95$.
- `alternative`: Við gefum skipunina `alternative="two.sided"` ef gagntilgátan er tvíhliða, `alternative="greater"` ef gagntilgátan er $\mu_1 - \mu_2 > \delta$ og `alternative="less"` ef gagntilgátan er $\mu_1 - \mu_2 < \delta$. Sjálfgefið er að hafa tvíhliða gagntilgátu.

Ályktanir um mismun meðaltala tveggja óháðra þýða

Segjum sem svo að við viljum bera saman fyrri púls nemenda eftir kynjum. Þar sem púls gögnin eru á löngu sniði gefum við skipunina:

```
t.test(puls$fyrriPuls~puls$kyn)

##
##  Welch Two Sample t-test
##
## data:  puls$fyrriPuls by puls$kyn
## t = 2.6808, df = 358.94, p-value = 0.007684
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.8000951 5.2065375
## sample estimates:
## mean in group Kona mean in group Karl
##           73.04082           70.03750
```

Ályktanir um mismun meðaltala paraðra mælinga

Þegar t-próf er framkvæmt fyrir mismun paraðra mælinga er skipunin `t.test()` mötuð með stillingunni:

- paired=TRUE

Annars er skipunin er hún mötuð á nákvæmlega sama hátt og þegar borin eru saman meðaltöl tveggja óháðra þýða.

Ályktanir um mismun meðaltala paraðra mælinga

Í púlsgögnunum liggur beint við að bera saman fyrri og seinni púls þeirra nemenda sem að hlupu í eina mínútu.
Byrjum á því að búa til tvær minni gagnatöflur, eina fyrir þá nemendur sem hlupu og aðra fyrir þá sem hlupu ekki.

```
pulshljop <- filter(puls, inngrip=='hljop')
pulskyrr <- filter(puls, inngrip=='sat_kyrr')
```

Ályktanir um mismun meðaltala paraðra mælinga

Könnum tilgátuna að púlsinn sé frábrugðinn fyrir og eftir krónukastið fyrir þá sem hlupu. Athugið að núna eru pöruðu mælingarnar tvær geymdar í tveimur dálkum og því eru gögnin á víðu sniði með því tilliti. Því mötum við skipunina á eftirfarandi hátt:

```
t.test(pulshljop$fyrriPuls, pulshljop$seinniPuls, paired=TRUE)


##
##  Paired t-test
##
## data:  pulshljop$fyrriPuls and pulshljop$seinniPuls
## t = -19.421, df = 179, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -28.40310 -23.16357
## sample estimates:
## mean of the differences
##                -25.78333
```

Hvert erum við komin...

Línulegt samband

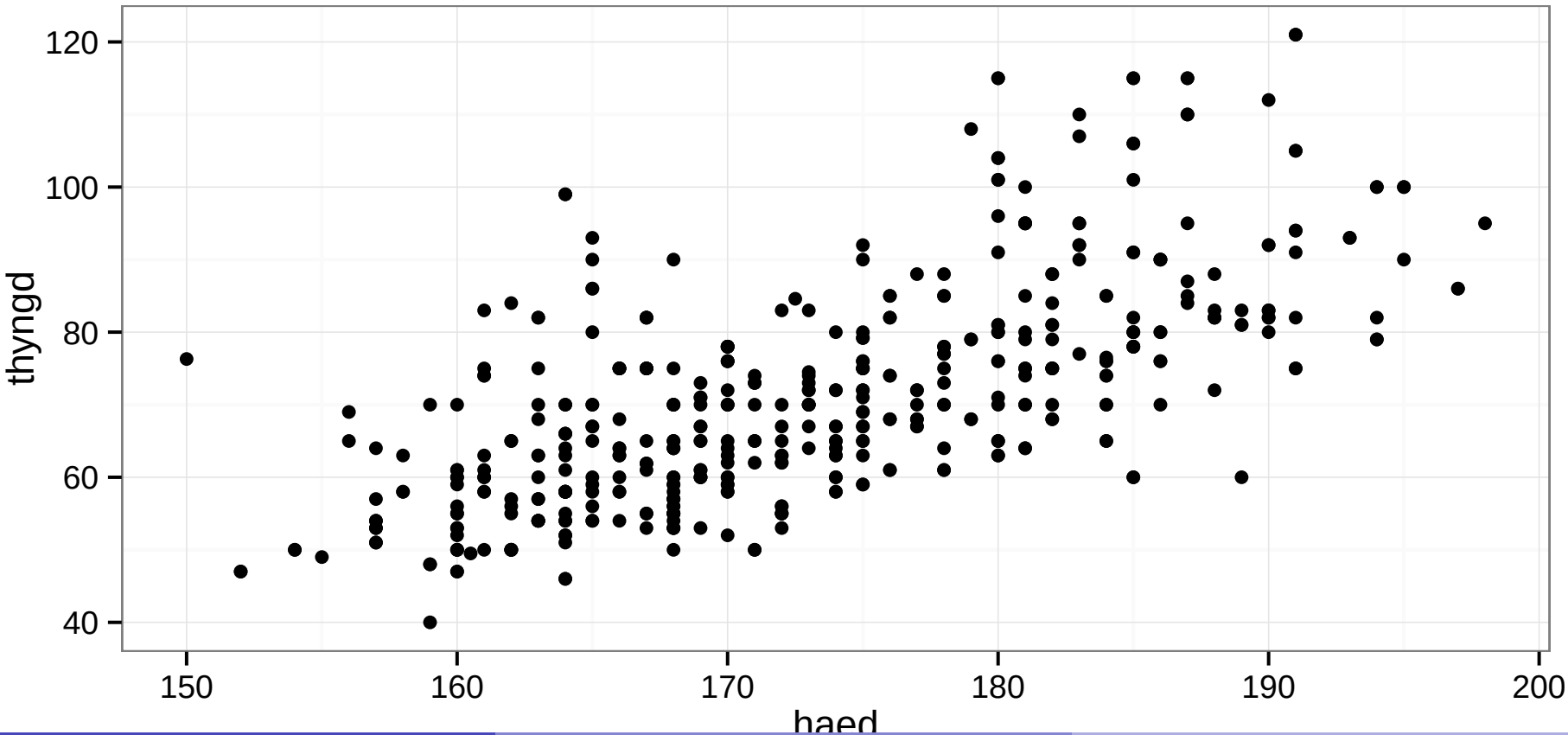
Línulegt samband

Við segjum að samband tveggja breyta sé *línulegt* (linear) ef nota má jöfnu beinnar línu til spá fyrir um gildi háðu breytunnar út frá gildi óháðu breytunnar.

Athugið að það geta verið margs konar aðrar gerðir af samböndum milli tveggja breyta. Til dæmis ef lýsa má sambandinu með fleygboga, veldisvísisfalli og svo framvegis. Þau sambönd eru einu orði nefnd ólínuleg og eru utan efni þessa námskeiðs.

Er línulegt samband milli hæðar og þyngdar?

```
library(ggplot2)
ggplot(puls, aes(x=haed, y=thyngd)) +
  geom_point() + theme_bw()
```



Línulegt líkan metið í R

Við metum línulegt aðhvarfsgreiningarlíkan með skipuninni `lm()`. Á vinstri hönd er svarbreytan (y-breytan) en skýribreyturnar á hægri hönd.

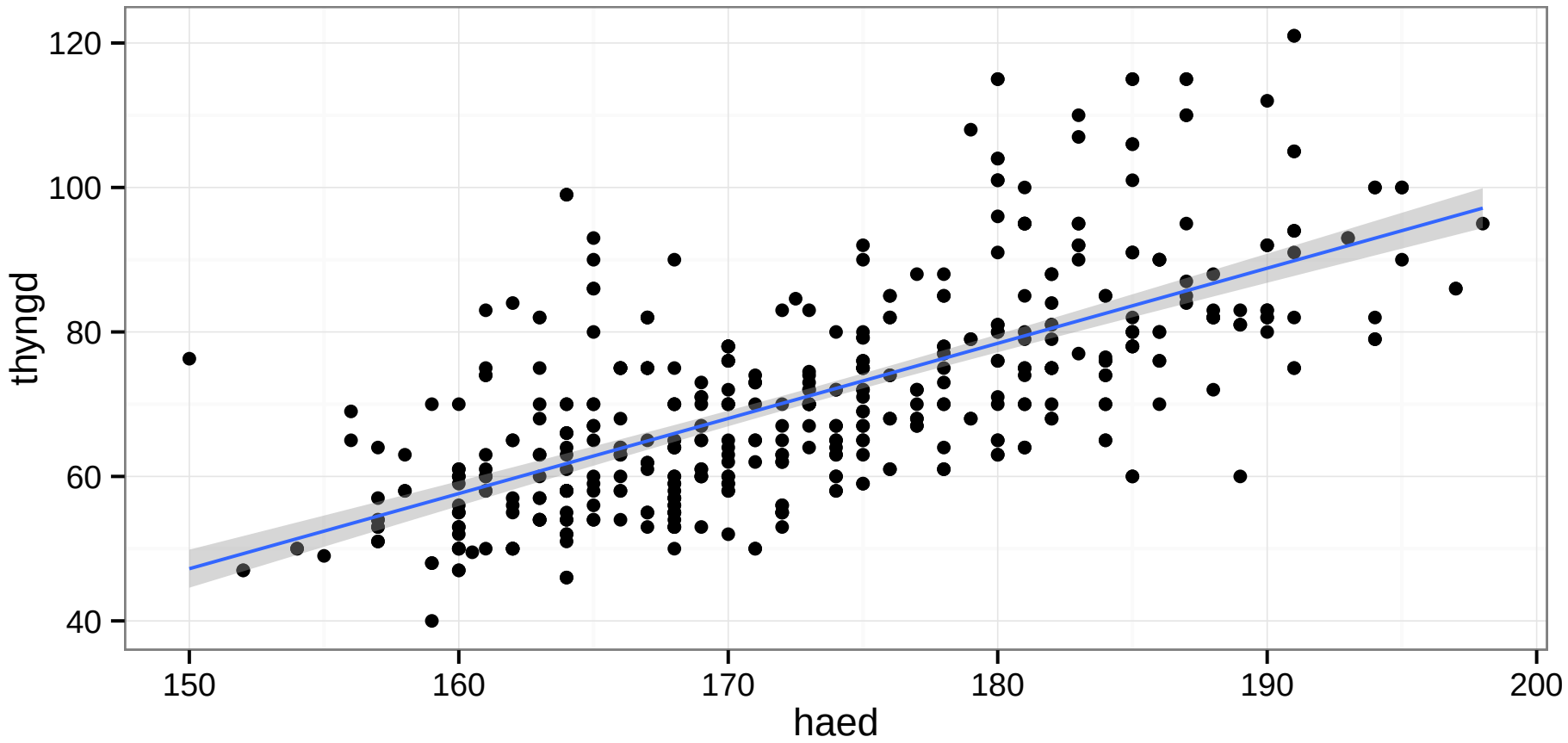
```
lm1 <- lm(thyngd~haed, data=puls)
lm1

##
## Call:
## lm(formula = thyngd ~ haed, data = puls)
##
## Coefficients:
## (Intercept)         haed
##     -108.75         1.04
```

Skurðpunkturinn er -108.754 en hallatalan 1.04.

Aðhvarfslínan komin á grafið.

```
library(ggplot2)
ggplot(puls, aes(x=haed, y=thyngd)) +
  geom_point() + theme_bw() + geom_smooth(method='lm')
```



Ályktanir um línulegt líkan í R - summary()

```
summary(lm1)


##
## Call:
## lm(formula = thyngd ~ haed, data = puls)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.781  -7.473  -1.984   5.306  37.215
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -108.75402    9.18021  -11.85   <2e-16 ***
## haed           1.03987    0.05288   19.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.17 on 458 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.4577, Adjusted R-squared:  0.4566
## F-statistic: 386.6 on 1 and 458 DF,  p-value: < 2.2e-16
```

Öryggisbil fyrir stuðla líkansins

Skipunin `confint.default` reiknar öryggisbil fyrir stuðla línulegs líkans.

```
confint.default(lm1)

##                    2.5 %     97.5 %
## (Intercept) -126.7468951 -90.761136
## haed           0.9362165   1.143522
```

Efri línan sýnir öryggismörkin fyrir skurðpunktinn.

Neðri línan sýnir öryggismörkin fyrir hallatöluna.

Spá um framtíðarmælingar og spábil

Skipunin `predict` reiknar spá og spábil fyrir gefin gildi á skýribreytum:

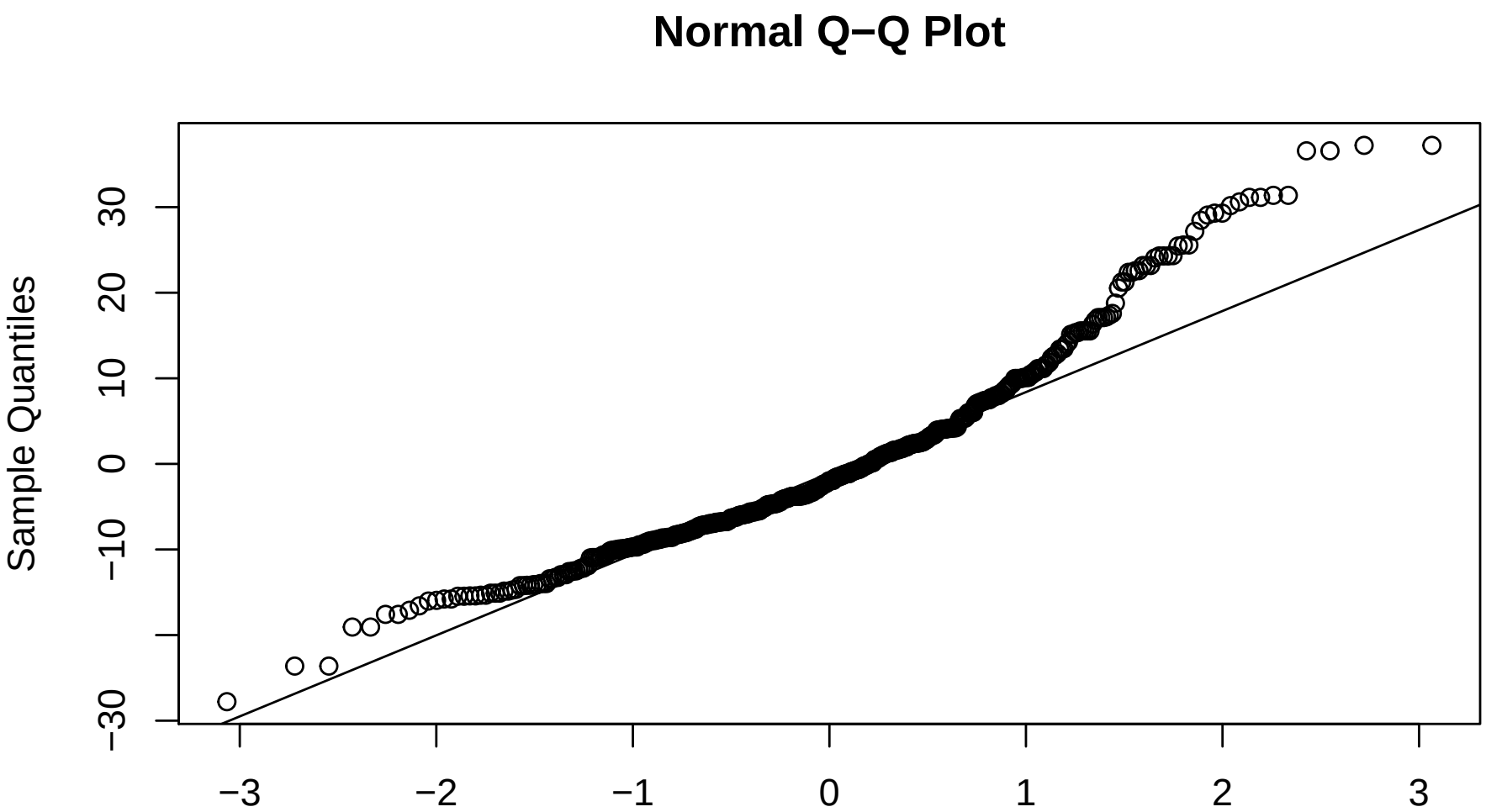
```
predict(lm1, data.frame(haed=175), interval='prediction')

##        fit     lwr      upr
## 1 73.22308 51.2456 95.20057
```

Fyrst sjáum við mat á spánni (`fit`) og þar næst neðra (`lwr`) og efra (`upr`) öryggismarkið.

Normaldreifingarrit: qqnorm() og qqline()

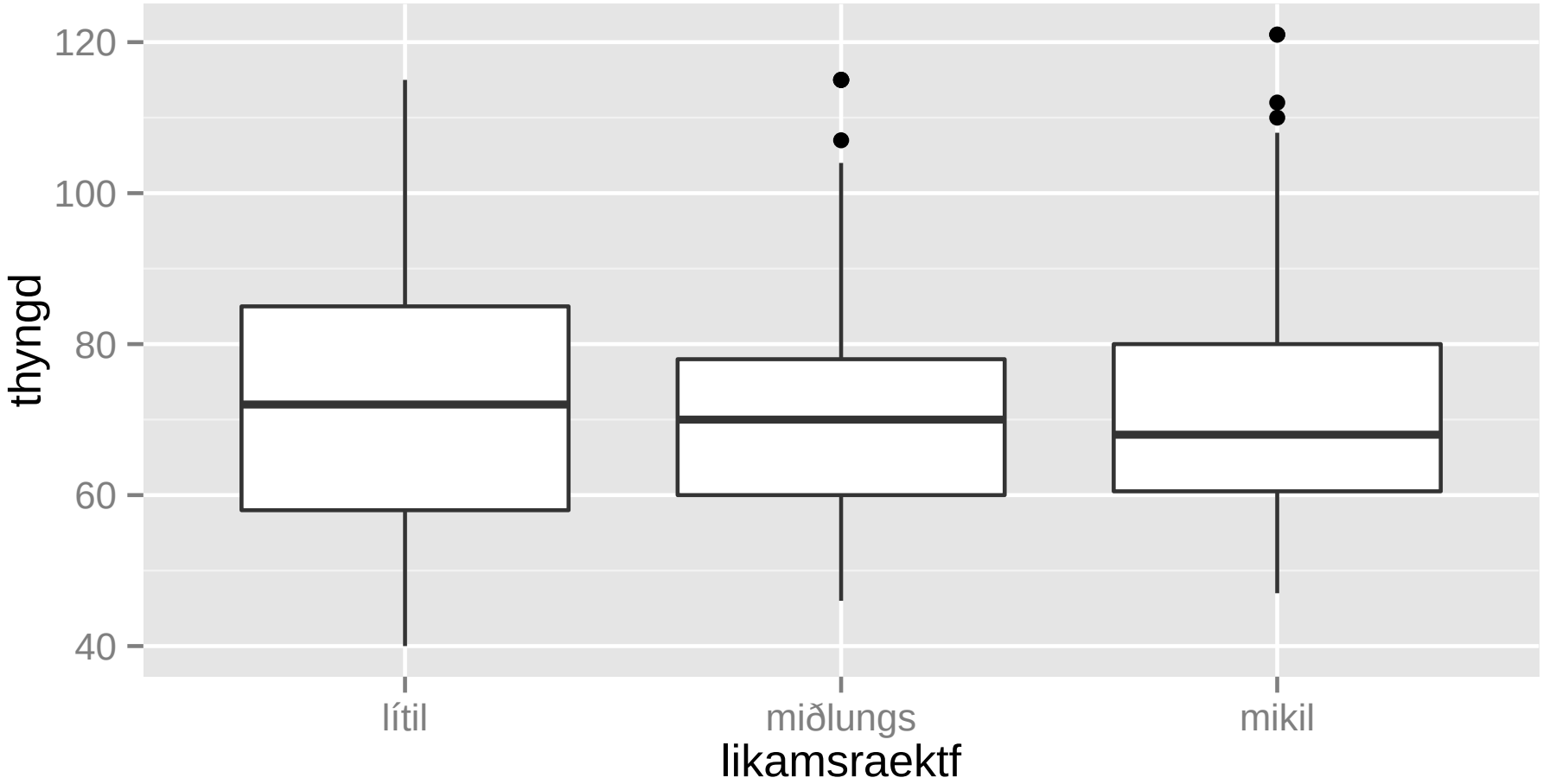
```
qqnorm(lm1$resid)
qqline(lm1$resid)
```



Hvert erum við komin...

Er munur á meðalþyngd eftir líkamsræktarástundun?

```
ggplot(subset(puls,!is.na(puls$likamsraekt)),
       aes(x=likamsraektf, y = thyngd)) + geom_boxplot()
```



```
fit.aov<-aov(thyngd~likamsraektf,data=puls)
summary(fit.aov)

##                Df Sum Sq Mean Sq F value Pr(>F)
## likamsraektf    2    691   345.7   1.517   0.22
## Residuals     455 103682   227.9
## 13 observations deleted due to missingness
```

Tukey próf

- Þegar við framkvæmum fervikagreiningu könnum við bara núlltilgátuna hvort öll meðaltölin séu eins
- Ef við höfnum núlltilgátunni þýðir það bara að amk eitt meðaltal sé frábrugðið hinum
- Við vitum ekki hvaða meðaltöl eru frábrugðin hverjum!
- Til þess að komast að því þarf að framkvæma eftiráþróf
- Yfirleitt á Tukey próf best við

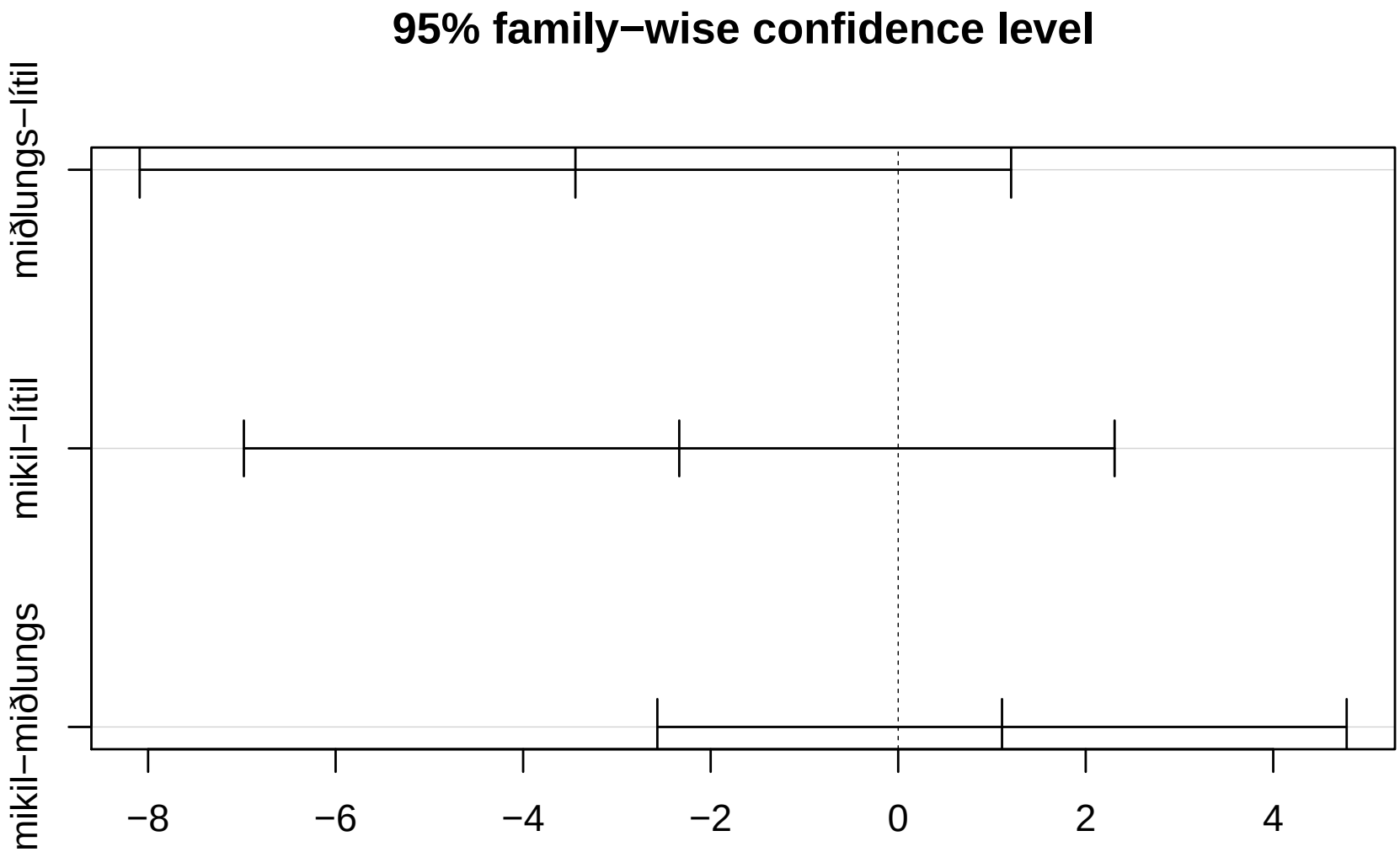
Tukey próf

```
TukeyHSD(fit.aov)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = thyngd ~ likamsraektf, data = puls)
##
## $likamsraektf
##                       diff       lwr      upr     p adj
## miðlungs-lítil -3.442568 -8.089773 1.204637 0.1908083
## mikil-lítil    -2.335294 -6.978600 2.308012 0.4641558
## mikil-miðlungs  1.107274 -2.568513 4.783060 0.7586952
```

Hvar liggur munurinn?

```
plot(TukeyHSD(fit.aov))
```



Hvert erum við komin...

Ákvörðunartré - spáð fyrir um gildi talnabreytu

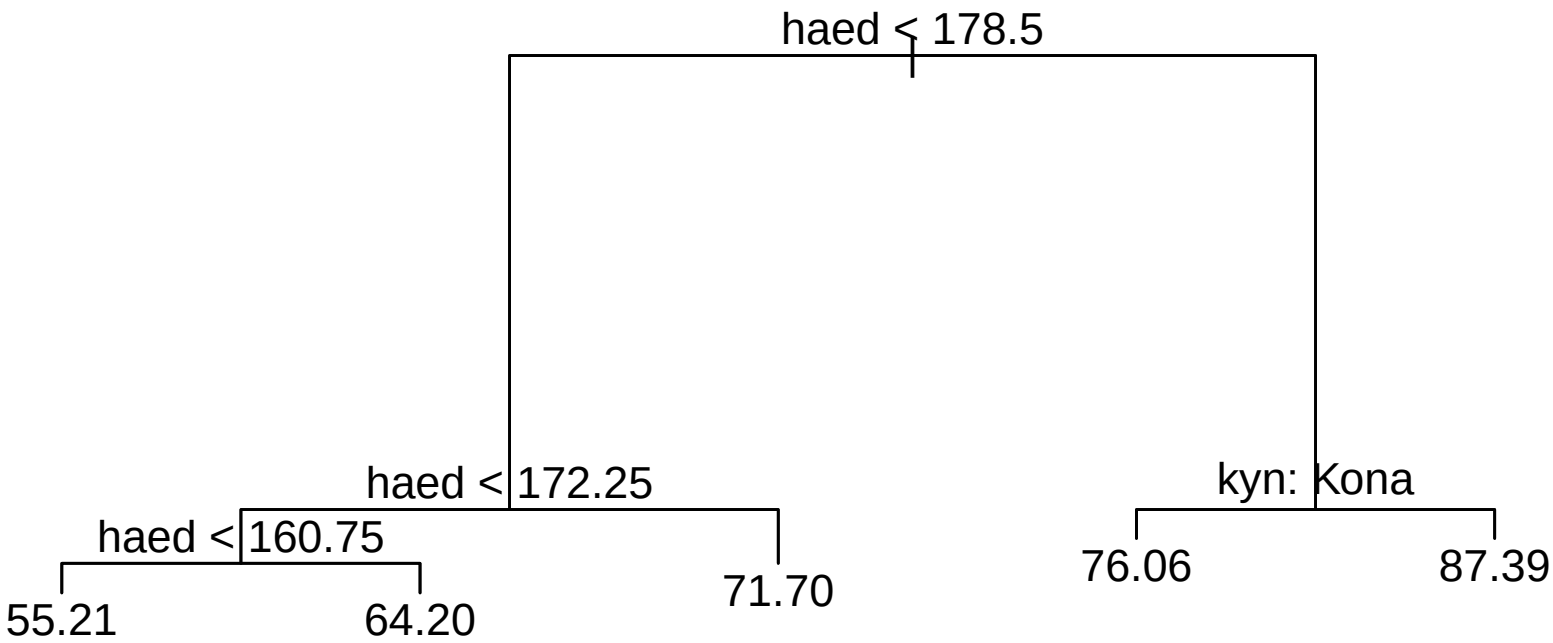
```
library(tree)
tree.1<-tree(thyngd~haed+kyn,data=puls)
summary(tree.1)

##
## Regression tree:
## tree(formula = thyngd ~ haed + kyn, data = puls)
## Number of terminal nodes:  5
## Residual mean deviance:  117 = 53240 / 455
## Distribution of residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -27.390  -7.388  -1.698   0.000   5.798  34.800
```

Hægt að snyrta (prune) tréð með `prune.misclass()`.

Ákvörðunartré - spáð fyrir um gildi talnabreytu

```
plot(tree.1)
text(tree.1,pretty=0)
```



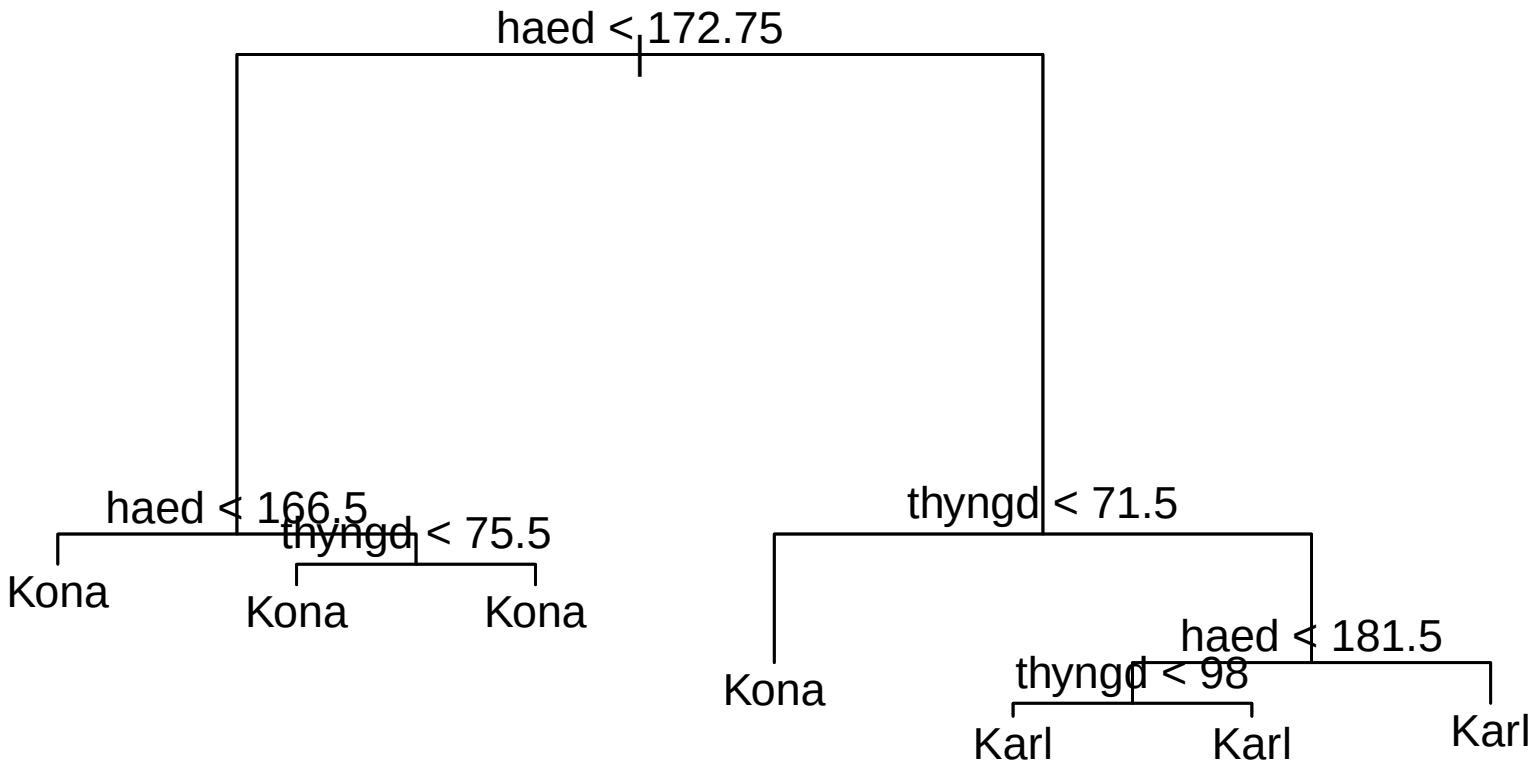
Ákvörðunartré - spáð fyrir um gildi flokkabreytu

```
library(tree)
tree.2<-tree(kyn~haed+thyngd,data=puls)
summary(tree.2)

##
## Classification tree:
## tree(formula = kyn ~ haed + thyngd, data = puls)
## Number of terminal nodes:  7
## Residual mean deviance:  0.5365 = 243 / 453
## Misclassification error rate: 0.1152 = 53 / 460
```

Ákvörðunartré - spáð fyrir um gildi flokkabreytu

```
plot(tree.2)
text(tree.2,pretty=2)
```



Bagging, random forests og boosting

- Helsti kostur "hefðbundna"ákvörunartrjáa eru hversu auðvelt er að túlka þau
- Ókostur: mikill breytileiki (high variance)
- Bagging, random forests og boosting: minni breytileiki
- Pakkar: `randomForest` og `bgm`

Hvert erum við komin...

Gögnin eins og þau koma „beint úr kúnni".

Náum í gögn: gengi evru, punds, dollara og gengisvísitala:

```
gengi1 <-read.table("gengi1.csv",header=T,sep=";")
head(gengi1)

##      Tstamp   EUR    GBP   USD    GVT
## 1  4.1.2000 73.51 117.38 71.61 110.10
## 2  5.1.2000 74.43 117.90 71.77 110.88
## 3  6.1.2000 74.32 118.06 71.75 110.73
## 4  7.1.2000 74.33 118.66 72.14 110.98
## 5 10.1.2000 74.07 118.44 72.33 110.85
## 6 11.1.2000 74.41 118.74 72.18 111.00
```

Dagsetningar eru sjálfkrafa vistaðar sem `factor`

```
str(gengi1)

## 'data.frame': 1621 obs. of  5 variables:
##  $ Tstamp: Factor w/ 1621 levels "10.10.2000","10.10.2001",.
##  $ EUR   : num  73.5 74.4 74.3 74.3 74.1 ...
##  $ GBP   : num  117 118 118 119 118 ...
##  $ USD   : num  71.6 71.8 71.8 72.1 72.3 ...
##  $ GVT   : num  110 111 111 111 111 ...
```

Við viljum vista `TSTAMP` sem tíma- og/eða dagsetningu.

as.Date skipunin

Við notum `as.Date` til að vista dagsetningar réttilega sem slíkar. Takið eftir `format` stillingunni. Hún er löguð að sniði gagnanna:

```
gengi1$dags <- as.Date(gengi1$Tstamp,format="%d.%m.%Y")
```

- Sé sniðið 1982/02/03 notum við format="%Y/%m/%d"
- Sé sniðið 03 02 82 notum við format="%d %m %y" o.s. frv.

Nánar um stillinguna format

`format` hefur stillingar fyrir nánast hverja einustu leið til að tákna dagsetningar. Helstu atriði eru:

- %Y eða %y - fjögurra eða tveggja stafa ár (2008 eða 08).
- %m - mánuður í tölustöfum
- %B eða %b - mánuður í bókstöfum (allt orðið eða skammstafað).
- %d - dagurinn

Einnig er t.d. hægt að vista tímasetningu með dagsetningu.

Skoðið `help(strftime)` fyrir fleiri stillingar á format.

Handhægar upplýsingar um dagsetningar

R hefur fullt af innbyggðum skipunum sem nota má til að fá handhægar upplýsingar

```
weekdays()    # vikudagur dags.
months()      # mánuður
quarters()    # ársfjórðungur
julian()      # fjöldi daga síðan áku. dags.
difftime()    # lengd milli tímasetninga
              # má stilla frá sek. upp í ár.
```

Unnið með dagsetningar úr SAS eða SPSS

Dagsetningar eru skráðar í SAS sem fjöldi daga síðan 1. janúar 1960. Dagurinn í dag væri því skráður sem 19471. Við finnum tilsvarandi dag með:

```
as.Date('1960-01-01') + 19471
```

Í SPSS eru dagsetningar vistaðar sem fjöldi sekúntna síðan á miðnætti 15. október 1582. Dagurinn í dag væri því skráður sem 13523070000. Við finnum réttan dag með:

```
as.Date('1582-10-15') + 13523070000
```

timeDate pakkinn

- Kemur frá Rmetrics hópnum.
- Inniheldur margar frábærar aðferðir til að vinna með dagsetningar.
- Getum skilgreint tímabelti, sem og kauphallir
- Er með innbyggðar dagsetningar á öllum helstu hátíðisdögum - ekki bara á vesturlöndum.

Dæmi um aðferðir í timeDate

```
library(timeDate)
dayOfWeek()                 #hvaða dagur vikunnar?
dayOfYear()                 #hvaða dagur ársins?
timeLastDayInMonth()        #síðasta dags. í þessum mánuði
timeFirstDayInQuarter()     #fyrsta dags. í þessum ársfjórðungi
Easter()                    #hvenær eru páskar?
holiday()                   #er helgidagur?
isBizday()                  #er viðskiptadagur?
```

lubridate pakkinn

- Útvíkkar timeDate.
- Inniheldur frábærar aðferðir til að vinna með dagsetningar á fágaðri hátt.
- Getum skilgreint tímabil á þægilegan hátt.
- Aðferðir fyrir tímasetningar.

Dæmi um föll í `lubridate` pakkanum:

```
library(lubridate)
as.period()      #skilgreina tímabil
floor_date()     #námunda niður dagsetningar
with_tz()        #fá tímasetningu í öðru tímabelti
leap_year()      #er hlaupár?
```

Pakkar fyrir tímaraðir

R státar af mörgum pökkum til að vinna með tímaraðir.

Þeir helstu eru:

- `ts`: Elsti pakkinn en með flestöllum hefðbundnum aðferðum.
- `xts` og `zoo`: "Nýrri og betri", mikið notaðir og fjölmikið í boði fyrir þá.
- `timeSeries`: Sá nýjasti og jafnframt sá sem við mælum með.
  Verður væntanlega arftaki hinna.
  Kemur frá Rmetrics teyminu.

timeSeries skipunin

Skipunin `timeSeries()` vistar gögn sem tímaraðir. Hún tekur fyrst inn gögnin, svo tímasetningarnar.

```
library(timeSeries)

## Loading required package:  timeDate

gengi1.ts <- timeSeries(gengi1[,-c(1,6)], gengi1$dags)
head(gengi1.ts)

## GMT
##                EUR    GBP   USD    GVT
## 2000-01-04 73.51 117.38 71.61 110.10
## 2000-01-05 74.43 117.90 71.77 110.88
## 2000-01-06 74.32 118.06 71.75 110.73
## 2000-01-07 74.33 118.66 72.14 110.98
## 2000-01-10 74.07 118.44 72.33 110.85
```

Einfalt að reikna

Við getum valið úr og reiknað með tímaraðir á nákvæmlega sama hátt og fylki.

```
gengi1.ts[100:104,c(1,4)]

## GMT
##                  EUR    GVT
## 2000-05-26 70.14 108.68
## 2000-05-29 70.51 108.73
## 2000-05-30 70.84 108.80
## 2000-05-31 70.52 108.62
## 2000-06-02 70.55 108.39
```

```
head(gengi1.ts$EUR + gengi1.ts$USD)

## [1] 145.12 146.20 146.07 146.47 146.40 146.59
```

Tímaraðir sameinaðar

Náum í gögn fyrir aðra gjaldmiðla:

```
gengi2<-read.table("gengi2.csv",header=T,sep=";")
gengi2$dags <- as.Date(gengi2$Tstamp,format="%d.%m.%Y")
gengi2.ts <- timeSeries(gengi2[,-c(1,5)], gengi2$dags)
head(gengi2)
```

```
##      Tstamp    SPX    CRY   TLT       dags
## 1 26.7.2002 852.84  168.3 82.51 2002-07-26
## 2 29.7.2002 898.96 166.87 81.42 2002-07-29
## 3 30.7.2002 902.78 169.45 81.52 2002-07-30
## 4 31.7.2002 911.62 169.78 82.53 2002-07-31
## 5  1.8.2002 884.66 168.18 83.00 2002-08-01
## 6  2.8.2002 864.24 168.84 83.85 2002-08-02
```

Tímaraðir sameinaðar

Tímaraðir sameinast eftir dagsetningum.

```
gengi3.ts<-cbind(gengi1.ts,gengi2.ts)
gengi3.ts[1:3,]

## GMT
##                EUR    GBP    USD    GVT     SPX   CRY   TLT
## 2000-01-04 73.51 117.38 71.61 110.1   <NA> <NA> <NA>
## 2000-01-05 74.43 117.9  71.77 110.88 <NA> <NA> <NA>
## 2000-01-06 74.32 118.06 71.75 110.73 <NA> <NA> <NA>

gengi3.ts[1500:1503,]

## GMT
##                EUR    GBP    USD    GVT     SPX      CRY    TLT
## 2005-11-29 74.68 109    63.21 104.9  1257.48 311.23  90.10
## 2005-11-30 74.63 109.25 63.35 104.91 1249.48 314.27  90.00
## 2005-12-01 74.42 109.35 63.16 104.62 1264.67 320.71  89.47
## 2005-12-02 74.54 110.03 63.66 105.11 1265.08 323.38  89.50
```

Ítarefni

- www.rmetrics.org
- https://www.rmetrics.org/ebooks-tseries
- http://cran.r-project.org/web/packages/timeDate/timeDate.pdf
- http://cran.r-project.org/web/packages/timeSeries/timeSeries.pdf

Hvert erum við komin...

Íslenskir stafir

- Makkar og Linux vélar: stafagerð (encoding) UTF-8
- Windows vélar: stafagerð (encoding) ISO-8859-1
- Ef þið fáið skipanaskrár (.R) eða aðrar skrár (.Rmd, .Rnw) og íslenskir stafir birtast ekki rétt passið að vista ekki skrána og farið í `File`, `Reopen with encoding` og veljið þá stafagerð sem notuð var þegar skráin var búin til. Vistið svo skrána og notið þá stafagerð sem á við ykkar stýrikerfi.
- Ef íslenskir stafir birtast ekki rétt á myndum inni í R og/eða þegar þið skoðið gagnatöflur keyrið þá eftirfarandi í keyrslugluggunum ykkar

```
system("defaults write org.R-project.R force.LANG en_US.UTF-8")
```

  Þetta þarf aðeins að gera einu sinni.

- Ef þið lengið í vandræðum með að lesa inn skrár með íslenskum stöfum með `read.table()` skoðið þá `encoding` stillinguna.