Þróun

Snæbjörn Pállsson & Áki Jarl Láruson

10 janúar, 2022

## R-kynning

### Um R-ið

R-ið er forrit sem hægt er að nota við tölfræðigreiningar, reikninga, til að teikna myndir, skrifa forrit og einnig við skýrslugerð. Forritið er ókeypis og hægt er að hlaða því niður frá: <http://www.r-project.org>, þar sem má einnig finna ýmsar gagnlegar upplýsingar um notkun Rsins. Veljið CRAN og “base” útgáfu fyrir það stýrikerfi sem þið eruð með á tölvunni.

Í “base” eru grunnskipanir til að keyra R-ið. Auk “base” eru til fjöldi pakka með mismunandi skipunum fyrir sérhæfðari vinnu sem hægt er að bæta við síðar en þið þurfið ekki núna t.d. til að greina gögn um samfélög lífvera eða þróunartré. Til að keyra R-ið er hægt að nota mismunandi forrit en hér er mælt með R-studio <http://www.rstudio.com>. Ókeypis útgáfan dugar vel.

Þegar þið eruð búin að setja upp bæði R-ið og R-studio þá getið þið klikkað á Rstudio-íkoninn og þá opnast þrír eða fjórir gluggar:

1. Ef þið klikkið á íkoninn efst í horninu til hægri (grænn depill með hvítum krossi) þá getið þið valið R-script sem má nota til að skrifa skipanir. Þær skipanir má senda í “Console” gluggann fyrir neðan með því að ýta á Run.
2. Í “Console” glugganum neðst til vinstri er einnig hægt að framkvæma skipanir á eftir “>” t.d. reikninga. Þar er einnig hægt að hlaða inn gögnum í R-ið eða búa þau til (sjá dæmi síðar). Þetta er vinnusvæðið (e. workspace) og það má vista það undir ólíkum nöfnum og geyma í mismunandi möppum eða sem verkefni (File -> New project sjá síðar) líkt og word-skjöl sem þið getið haldið að vinna áfram með síðar. Þá getið þið farið inn í möppuna og klikkað á R-íkoninn til að opna Rstudio sem þið voruð að vinna með síðast.
3. Efri glugginn til hægri sýnir hvaða gögn eru til staðar í því vinnusvæði sem er opið (environment), hvaða skipanir hafa verið gefnar (History) og þar er flipi til að lesa inn gögn (líka hægt að nota R-skipun t.d. read.csv).
4. Neðri glugginn til hægri er með flipa þar sem finna má yfirlit yfir skrár í sömu möppu og þið eruð að vinna í, myndir sem þið hafið teiknað í R-inu, pakka sem hægt er að ná í á heimasíðu R-sins, hjálp fyrir R-föll (functions), og gögn sem fylgja base eða þeim pökkum sem þið hafið tengt við vinnusvæðið.

### R sem reiknivél:

# R les ekki texta á eftir #  
2+2 # ýtið á enter

## [1] 4

2\*2

## [1] 4

2^2

## [1] 4

### Föll

Ýmis föll eru skilgreind í R og eru þau notuð með því að nota () á eftir nafni fallsins. Inn í svigann er hægt að setja gögn og ákveðnar upplýsingar (argument) sem hefur hefur áhrif á hvað fallið gerir.

a = c(2,4,2,3) # c-fallið bindur saman tölurnar 2, 4, 2, og 3 í vektor  
a # enter

## [1] 2 4 2 3

mean(a) #gefur meðatal af gildum í a

## [1] 2.75

help(mean) # hjálparsíða fylgir öllum skipunum  
# velja má ákveðin gildi með því að setja ákv. skilyrði innan []:  
a[1] # fyrsta gildið í "a"

## [1] 2

a[1:3] # hvað gerir 1:3? Prófið

## [1] 2 4 2

1:3 #enter

## [1] 1 2 3

# hægt er að framkvæma reikninga á vektorum t.d.  
3\*a

## [1] 6 12 6 9

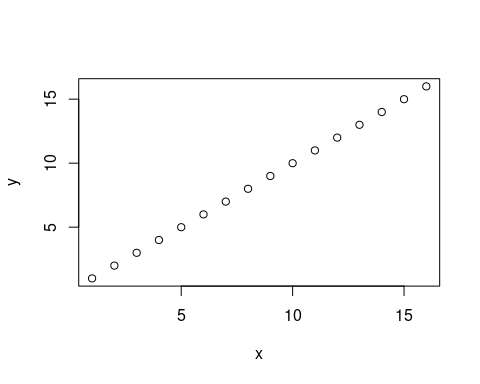
a+a

## [1] 4 8 4 6

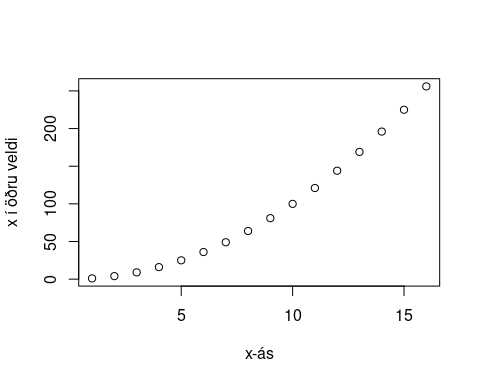
### Myndir

Til að teikna myndir eru tvennskonar skipanir, háskipanir sem teikna nýjar myndir og lágskipanir sem bæta upplýsingum inn á myndir:

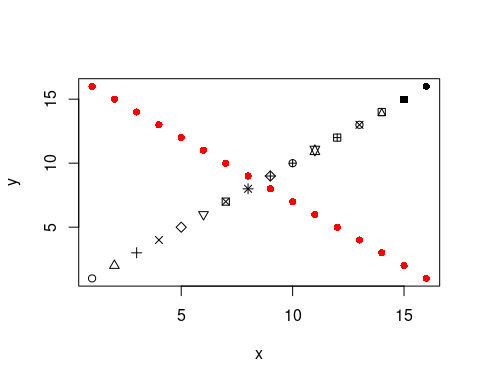
x <- 1:16 # hægt er að nota <- eða =, líka ->  
y = x # enter  
plot(x,y)



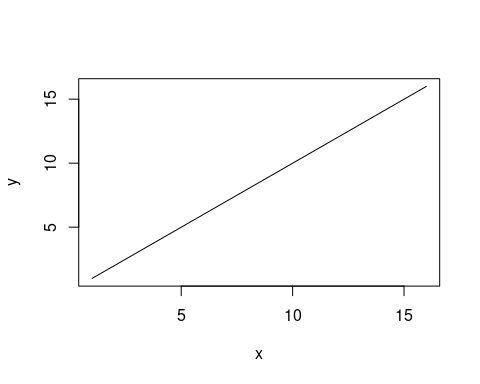
plot(x,y^2,xlab="x-ás",ylab="x í öðru veldi")



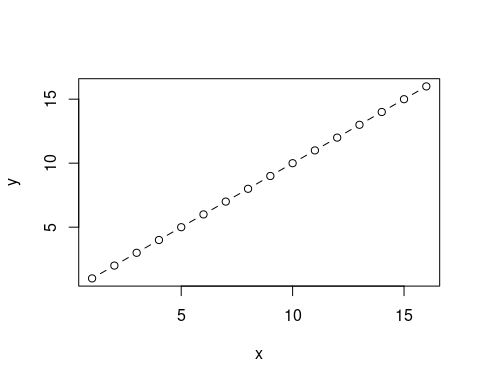
#ath hvað gerði xlab og ylab  
plot(x,y,pch=1:16) # pch e. plot character gefur mismunandi tákn   
points(x,y[16:1],pch=16,col="red") # öfug röð á y með lit



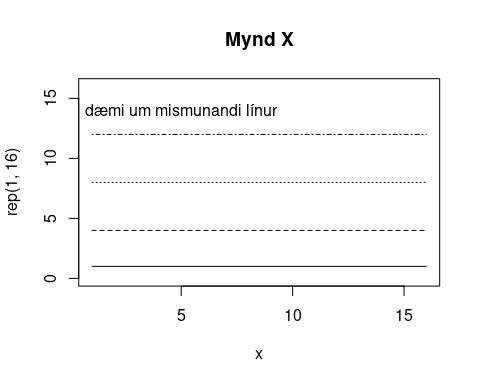
plot(x,y,type="l") # "l" fyrir línu



plot(x,y,type="b") # bæði punktar og lína



plot(x,rep(1,16),type="l",ylim=c(0,16)) # lty = line type, ylim: y limits  
lines(x, rep(4,16),lty=2)  
lines(x, rep(8,16),lty=3)  
lines(x, rep(12,16),lty=4)  
text(5,14,"dæmi um mismunandi línur")  
title("Mynd X")



#hvað gerir rep? prófið að gera skipunina rep(1,16)  
rep(1,16)

## [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

### Lesa gögn inn í R og vinna með töflur

Best er að vista gögn sem tengjast ákveðinni vinnu eða námskeiði í einni möppu. Veljið eða búið til möppu sem heitir Líffræði, veljið File->New project og gefið því nafnið Rkynning í möppunni Líffræði.

Í Líffræðimöppunni er nú komin mappa sem heitir Rkynning og er með R-íkon sem hægt er að klikka á til að ræsa Rstudio. Í þessa möppu vistast vinnusvæðið sem .RData og einnig skráin .Rhistory með öllum skipunum. Skrifið q() til að hætta og veljið að vista eða notið File-Quit Session og vistið. Opnið Rstudio-ið með því að klikka á R-íkoninn í Rkynning.

Búið til gagnaskrá með því að opna excel. Notum gögn um heila (g) og líkamsstærð (kg) nokkurra spendýra. Skrifíð eftirfarandi gögn í þrjá dálka:

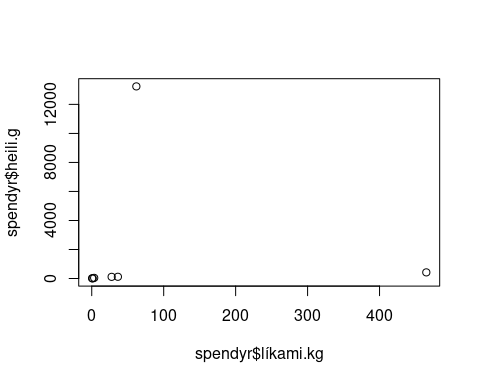
#tegund líkami.kg heili.g  
#Refur 3,385 44,5  
#Uglu-api 0,48 15,5  
#Bifur 1,35 8,1  
#Kýr 465 423  
#Úlfur 36,33 119,5  
#Geit 27,66 115  
#Maður 62 13239

Efsta línan eru dálkanöfnum, í R er vísað í þau sem “header”, sjá síðar. Til að lesa þessa skrá inn í R þarf að vista hana á csv-formati, sem er einfalt textaskjal sem má skoða í hvaða ritli sem er t.d. notepad. Veljið save file og csv format, hægt er að velja aðgreiningu milli dálka sem bil, kommu eða semikommu. Vistið skrána í möppunni Rkynning sem spendyr.csv, annars þurfið þið að gefa slóðina til að lesa inn skrána í R-inu.

spendyr=read.csv("spend.csv",sep=" ")   
#ath hjálp fyrir read.csv  
  
names(spendyr)

## [1] "Tegund" "líkami.kg" "heili.g"

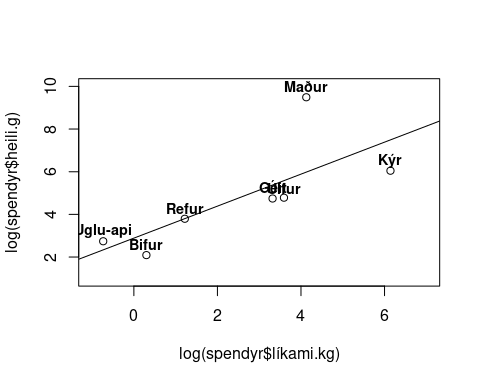
plot(spendyr$líkami.kg,spendyr$heili.g)



# betra væri að umbreyta gögnunum t.d. með logarithma (log)  
plot(log(spendyr$líkami.kg),log(spendyr$heili.g),xlim=c(-1,7),ylim=c(1,10))   
  
# finnum bestu beinu línu, notum fallið lm  
heili.lm=lm(log(heili.g)~log(líkami.kg),data=spendyr)  
  
heili.lm # hér má finna skurðpunkt (a) og hallatölu línunnar (b)

##   
## Call:  
## lm(formula = log(heili.g) ~ log(líkami.kg), data = spendyr)  
##   
## Coefficients:  
## (Intercept) log(líkami.kg)   
## 2.8888 0.7499

abline(heili.lm) # teiknum línuna y = a + bx  
  
#hvaða tegundir sýna mest frávik frá línunni?  
#notið identify(log(spendyr$líkami.kg),log(spendyr$heili.g),spendyr$Tegund)  
# klikkið á punktana og síðan á "Finish" takkann í horninu til # hægri fyrir ofan myndina  
#   
#Eða þá notið:  
text((log(spendyr$heili.g)+0.5)~log(spendyr$líkami.kg), labels=spendyr$Tegund,data=spendyr, cex=0.9, font=2)



### Erfðafræði

1. lögmál Mendels - 1 lókus tvö allel 0 og 1.

Arfgerd = c(0,1)  
#Söfnum kynfrumu:  
sample(Arfgerd,size=1)

## [1] 1

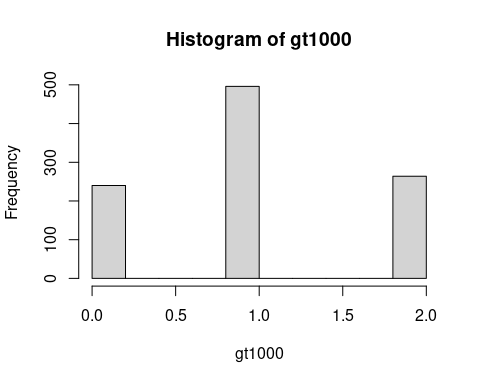
# æxlum saman tveimur arfblendnum einstaklingum  
sample(Arfgerd,r=T,size=2) #replace þýðir að sýnataka er með replacement #(þe hefur ekki áhrif á næstu sýnatöku)

## [1] 1 0

#og leggjum saman (afhverju)  
sum(sample(Arfgerd,r=T,size=2))

## [1] 0

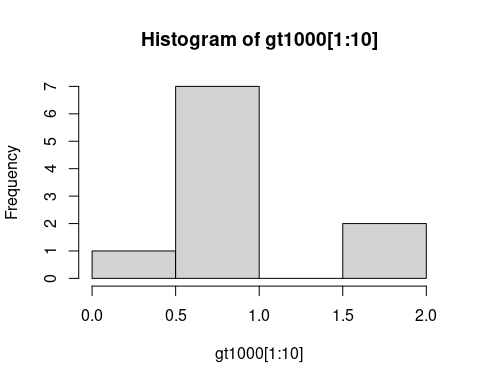
#Endurtökum fyrir 1000 æxlanir. Búum til vektor gt1000 og notum for() fallið:  
gt1000=1:1000  
for(i in 1:1000) gt1000[i] = sum(sample(Arfgerd,r=T,size=2))   
hist(gt1000)



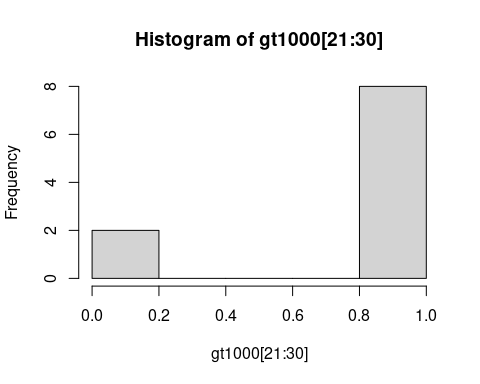
# Hvernig getum við reiknað væntanlegu gildin?   
# Hér getum við notað reglur um líkur á óháðum og aðskildum atburðum þe.P (A og B) = P(A)\*P(B) og P(A eða B) = P(A)+ P(B).   
# Líkurnar á að fá 00, 01 og 11 eru þá 0.25, 0.5 og 0.25 þar sem P(0)=P(1)=0.5.   
# Afhverju eru líkurnar á 01 = 0.5? Hvernig ber þessu saman við séða fjöldann? Hér þarf að margfalda með heildarfjölda.

### Teiknið fyrir fyrstu 10 og fyrstu 1000

hist(gt1000[1:10])



# Hvernig ber fyrstu 10 og fyrstu 1000 saman?   
# Prófið líka   
hist(gt1000[11:20]) ; hist(gt1000[21:30])



Ef við gerum ráð fyrir tilviljunarkenndri æxlun hjá hópi einstaklinga getum við reiknað arfgerðirnar fyrir hvaða hlutfallslega tíðni sem er í hópnum, þ.e. P(0) getur verið hvaða tala sem er frá 0:1. Líkurnar á að fá 00, 01 og 11 ef p eru líkunar á 0 og q líkurnar á 1 eru þá p^2, 2pq og g^2. Byrjum á að safna arfgerðunum og reiknum svo þetta út.

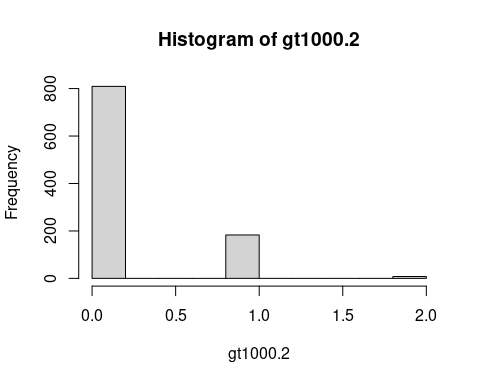
stofn=c(rep(0,90),rep(1,10))  
stofn # hvað er þetta?

## [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1

sample(stofn, 2, r=T)

## [1] 0 0

gt1000.2=1:1000  
for(i in 1:1000) gt1000.2[i] = sum(sample(stofn,r=T,size=2))   
hist(gt1000.2)



Hvað ef það eru margir lókusar t.d. 5? Gerum ráð fyrir að þeir erfist óháð hver öðrum (annað lögmál Mendels). Ef tveir einstaklingar æxlast saman og þeir eru arfblendnir fyrir sömu tvö allelin á 5 lókusum þá gæti ein kynfruma verið:

sample(Arfgerd,r=T,size=5)

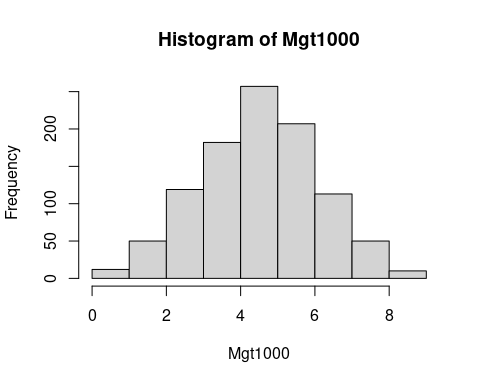
## [1] 1 1 1 0 0

Hvað mætti búast við mörgum ólíkum kynfrumum? En ef við værum með 100 gen og 10 afkvæmi (“sample(Arfgerd,r=T,size=10)”)? Ef við gerum ráð fyrir að áhrif hvers gens leggist saman og myndi eina svipgerð (e. phenotype) t.d. hæð má leggja saman áhrifin eins og áður t.d.

sum(sample(Arfgerd,r=T,size=10))

## [1] 2

Mgt1000=1:1000  
for(i in 1:1000) Mgt1000[i] = sum(sample(Arfgerd,r=T,size=10))  
hist(Mgt1000)



Athugið hér eru eingöngu arfgerðaráhrif. Ef við gerum ráð fyrir umhverfisáhrifum á hæðina mætti bæta þeim við t.d. Svipfar1000=Mgt1000+rnorm(1000). Hvað gerir rnorm? hist(Svipfar1000)

### Skoðið gildin.

R Markdown Til að skrifa skýrslur og halda utan um gagnavinnslu er gott að nota “R Markdown”, sem má nota til að búa til HTML, PDF og MS Word skrár (<http://rmarkdown.rstudio.com>). Klikkið á græna depilinn í horninu efst til hægri og veljið R Markdown. Nú opnast sýnidæmi um hvernig hægt er að búa til skýrslu. Rskipanir eru gefnar í innan þriggja komma t.d. utan um summary(cars), en cars er gagnasett í R-base og má skoða nánar með help(cars). Einsog segir í textanum á ensku þá getið þið núna ýtt á “Knit” flipann, og “prjónað” Rskipanirnar og textann saman í eina skýrslu. Prófið að breyta textanum og R skipununum og ýtið á Knit. Ef þið hafið valið Word þá opnast word skjal með texta og þeim myndum og reikningum sem þið gerðuð í R-inu.