Tölvuverkefni 2

Líkindareikningur og tölfræði: STÆ203G, HAG206G, MAS201F

Í þessu verkefni munið þið vinna með úrtaksdreifingu meðaltals, lögmál mikils fjölda og höfuðmarkgildisregluna sem stundum er nefnd höfuðsetning tölfræðinnar (e. central limit theorem). Hún segir að ef $X_1, ...X_n$ eru einsdreifðar og óháðar slembistærðir með $E[X_i] = \mu$ og $VAR[X_i] = \sigma^2$ og n er nægjanlega stórt má nálga dreifingu

$$\bar{X} = \frac{X_1 + X_2 + ... + X_n}{n}$$

með normaldreifingu með meðaltal μ og dreifni σ^2/n.

Þið munið vinna með sama gagnasafn og í Tölvuverkefni 1 en það er hluti þess gagnasafns sem notað var við endurmat fasteignamats íbúðarhúsnæðis 2017. Þið getið lesið ykkur til um gagnasafnið hér:

https://notendur.hi.is/ahj/Fasteignamat2017.pdf

Töflu yfir matssvæði er að finna á síðu 59.

Gagnaskráin sem þið eigið að vinna með má hlaða niður af Uglu (úr möppunni R-tölvuverkefni) en hana er einnig að finna hér:

https://notendur.hi.is/~ahj/husnaedi.txt

Þið eigið að vinna verkefnið í .Rmd skrá en skila .html skrá (LaTeX fólk má vinna í .Rnw skrá og skila .pdf skrá). .html skráin verður til í sömu möppu og .Rmd skráin þegar þið prjónið. Þið megið vinna verkefnið tvö og tvö saman en ekki í stærri hópum. Merkja þarf lausnir með nafni og HÍ-notendanafni þeirra sem unnu lausnina. Hlaða skal .html skrá inn á Uglu ekki síðar en föstudaginn 22. febrúar klukkan 23:59. Ekki verður tekið við lausnum eftir þennan tíma.

Frágangur gildir 10% af einkunn fyrir verkefni (merkja verkefni með nafni og notendanafni, merkingar á ásum, uppsetning, . . .).

0.0.0.1 a)

Lesið gagnaskrána inn með read.table() eða read.csv() skipun og geymið í hlut sem ber upphafsstafi þeirra sem verkefnið vinna (ef Gréta Halldórsdóttir og Sigurður Jónsson eru að vinna saman skulu þau nefna hlutinn gs en ef Atli Pétursson er að vinna einn skal hann nefna hlutinn sinn ap). Takið eftir mata má read.table() með URL-i.

Í þessum lið getur þurft dálítið möndl til að ná íslenskunni réttri. Veljið eina eða tvær neðangreindra skipana eftir því sem við á (takið # framan af því sem á að keyra):

```
#Sys.setlocale("LC_ALL", "is_IS.UTF-8") # ef með þarf - þ.e. ef íslenskan birtist ekki rétt
#bh<-read.csv("http://uc-media.rhi.hi.is/tmp/fdytk/husnaedi.txt",sep=";",fileEncoding = "UTF8")
bh<-read.table("https://notendur.hi.is/~ahj/husnaedi.txt", sep=";", header=T) # hugsanlega dugar þetta
```

Hver hópur á að vinna með íbúðareignir úr einu af erftirfarandi hverfum:

- Hlíðar (matssvæði 80)

- Grafarvogur: Rimar, Engi, Víkur, Borgir (matssvæði 130)
- Hólar, Berg (matssvæði 160)
- Árbær (matssvæði 200)

Til að ákvarða hvaða hverfi þið eigið að vinna með skulið þið nota kóðann hér að neðan. Nemandinn sem er á undan í stafrófinu setur inn fæðingardaginn sinn í set.seed() skipunina. Í kóðanum hér neðan er einstaklingurinn fæddur 3. maí.

```
set.seed(0305)
(hverfi<-sample(c(80,130,160,200),1))
```

```
## [1] 200
```

Þessi hópur ætti því að vinna með íbúðaeignir í Árbæ.

Notið svo filter skipunina til að velja úr þær línur úr stóru gagnatöflunni sem innihalda aðeins íbúðareignir úr viðeigandi hverfi (þið þurfið að velja aðrar gerðir eigna frá). **Munið að þið þurfið að keyra library(dplyr) áður en þið keyrið filter() skipunina ykkar**. Þið skuluð yfirskrifa stóru gagnatöfluna (þ.e. notið sama nafn).

0.0.0.2 b)

Teiknið stuðlarit af verði íbúðanna - munið að merkja ásana rétt. Tilgreinið meðalverð og dreifni verðs íbúðanna í samfelldu máli (ekki harðkóða gildin heldur látið R reikna þau út fyrir ykkur).

0.0.0.3 c)

Við lítum nú svo á að við höfum mælingar á öllu þýðinu, þ.e. íbúðirnar í gagnatöflunni ykkar eru þýðið og eiginleikinn sem við höfum mælt er verð þeirra. Við þekkjum því μ og σ^2.

Takið nú 10000 sinnum úrtak af stærð $n = 10$ úr þýðinu, reiknið meðaltal úrtakanna og geymið í vigri sem heitir staerd1. Þessi vigur á að innihalda 10000 meðaltöl. Endurtakið leikinn fyrir $n = 30$, $n = 100$ og $n = 200$ og nefnið vigrana staerd2, staerd3 og staerd4. Notið replicate() aðferðina til að auðvelda ykkur verkið.

0.0.0.4 d)

Í þessum lið eigið þið að teikna stuðlarit af staerd1, staerd2, staerd3 og staerd4. Til þessa að gera það mæli ég með að þið búið til gagnatöflu sem inniheldur vigrana fjóra. Það má gera með:

```
hmr.bh<-data.frame(staerd1, staerd2, staerd3, staerd4)
```

Þessi gagnatafla er á svo kölluðu víðu sniði (wide format) en áður en við getum teiknað stuðlaritin þurfum við að koma henni á langt snið (long format). Það má gera á auðveldan hátt með gather() aðferðinni en hún tilheyrir tidyr pakkanum. Þið þurfið að byrja á að hlaða honum niður en að því loknu getið þið notað eftirfarandi kóða til að koma gögnunum á langt snið:

```
library(tidyr)
hmr.bh.long<-gather(hmr.bh, staerd, medaltal)
```

Skoðið gagnatöfluna sem verður til áður en lengra er haldið og gangið úr skugga um að allt sé með felldu.

Teiknið nú stuðlarit af staerd1, staerd2, staerd3 og staerd4. Auðvelt er fá stuðlaritin fjögur með að nota facet_wrap() aðferðina, sjá t.d. hér: http://ggplot2.tidyverse.org/reference/facet_wrap.html. Skalinn á x-ásnum á að vera sá sami (sjálfgefna skillingin í facet_wrap() aðferðinni).

0.0.0.5 e)

Teiknið aftur stuðlarit af staerd1, staerd2, staerd3 og staerd4 en í þetta skiptið á skalinn á x-ásnum ekki að vera sá sami (skoðið scales viðfangið/stillinguna í facet_wrap() aðferðinni).

0.0.0.6 f)

Reiknið meðaltal og dreifni meðaltalanna sem fengust í lið c) hér fyrir framan fyrir $n = 10,\ n = 30,\ n = 100$ og $n = 200$ og sýnið niðurstöðuna í töflu.

0.0.0.7 g)

Lýsið stuttlega með eigin orðum því sem þið sjáið á myndunum úr lið d) og e) og niðurstöðum ykkar úr lið f) og hvernig það tengist höfuðmarkgildisreglunni.