Exercise 5

Peter Heilbo Ratgen

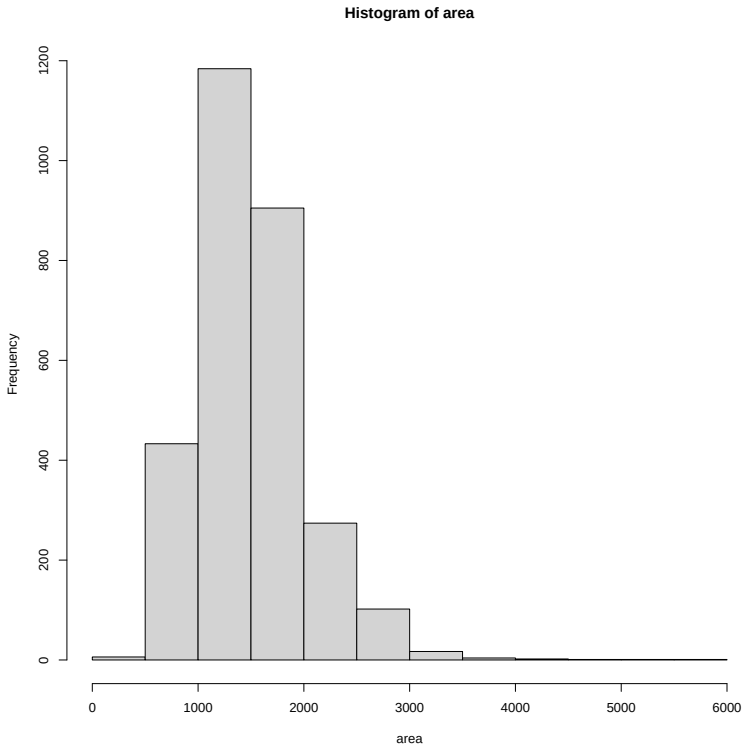
3. januar 2021

1 Data

Vi laver en summering af variablen `ames$Gr.Liv.Area`.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
334	1126	1442	1500	1743	5642

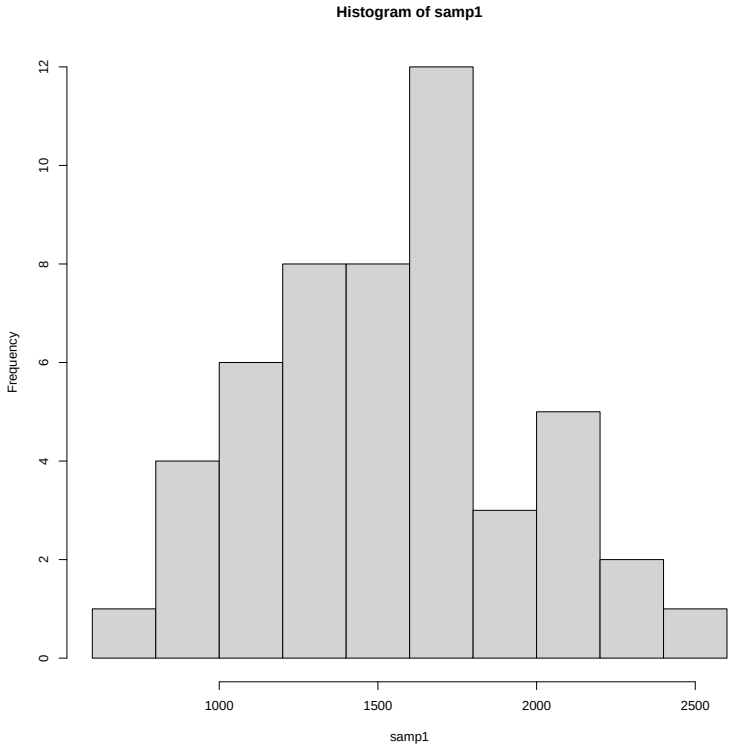
Vi laver også et histogram over denne data.



Her kan vi se at fordelingen er højreskæv, samt at den har den karakteristiske klokkeform som kendetegner normalfordelingen.

2 Den ukendte prøveudtagningsfordeling

Vi tager stikprøve med `samp1 <- sample(area,50)`. Her får vi 50 tilfældige værdier.



Fordelingen ligner tilnærmelsesvist en normalfordeling. Fordelingen kan ikke betegnes som værende højreskæv eller venstreskæv, men derimod symmetrisk.

Vi tager et gennemsnit af `samp1`, som en approksimering af gennemsnittet for den generelle population. Dette gennemsnit er 1520 sq ft, hvor gennemsnittet for den generelle population er 1499 sq ft.

Vi tager endnu en stikprøve fra populationen.

```
@> samp2 <- sample(area,50)
@> mean(samp2)
[1] 1489.94
```

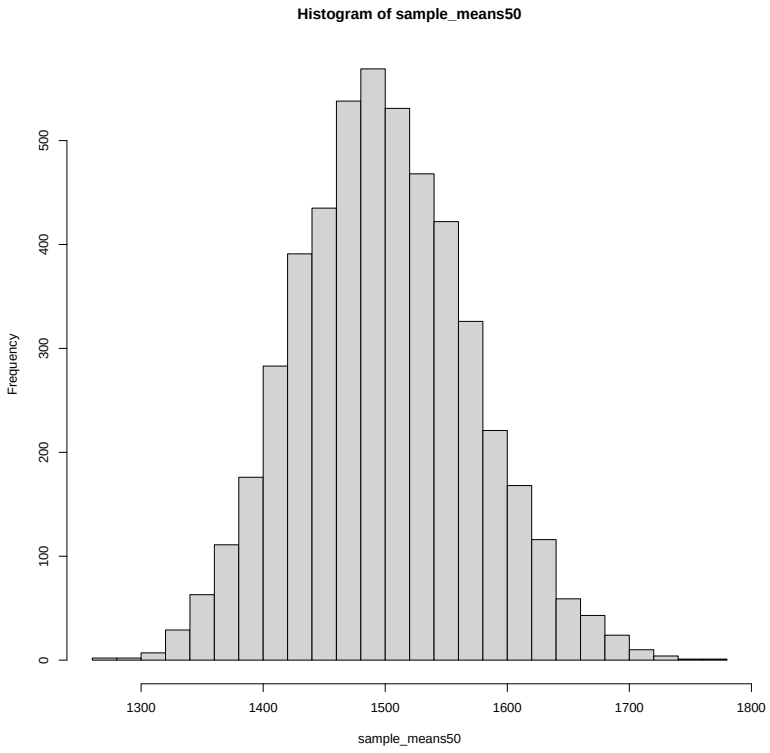
Denne stikprøve ligger heller ikke langt fra gennemsnit for den generelle population, den ligger heller ikke langt fra `samp1`. Den største prøve vil give et mest nøjagtige skøn, dette er dog ikke at foretrække, da fordelene i at få 1000 stikprøver frem for 50 stikprøver, ikke giver en meget bedre model.

2.1 5000 stikprøver

Vi genererer 5000 stikprøver og tager gennemsnittet.

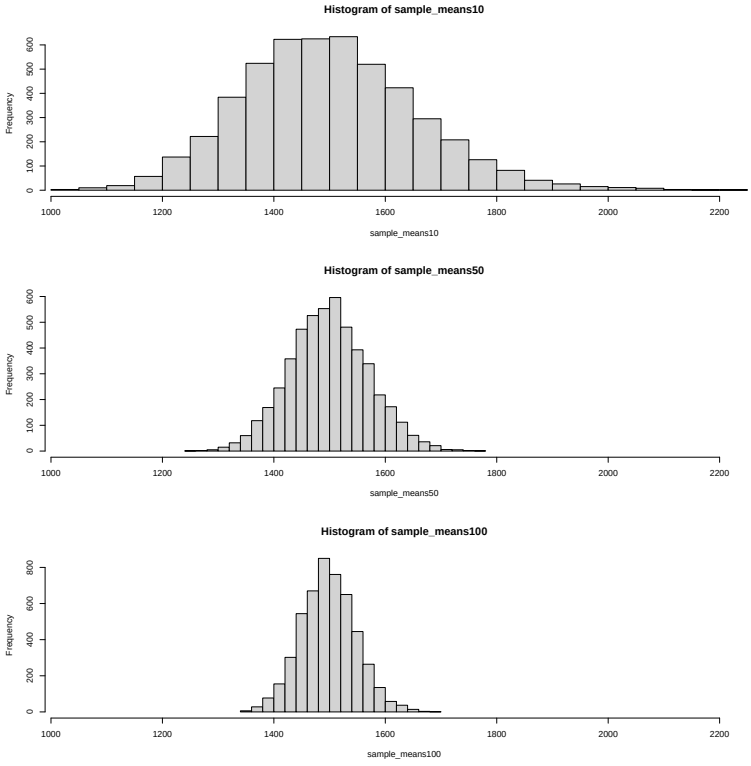
```
sample_means10 <- rep(NA, 5000)
sample_means50 <- rep(NA, 5000)
sample_means100 <- rep(NA, 5000)

for(i in 1:5000) {
  samp <- sample(area,10)
```

Der er 5000 elementer i `sample_means50`.



Gennesnittet for de 5000 prøver er

```
@> mean(sample_means50)
[1] 1498.804
```



Jo større stikprøvestørrelsen bliver, jo tættere bliver spredningen omkring middelværdien for den generelle population. Spredningen af data bliver mindre, siden vi har flere datapunkter med per værdi.

Det vil sige at en værdi i means_100 generelt set bedre repræsentere end en vilkårlig værdi i means_10.

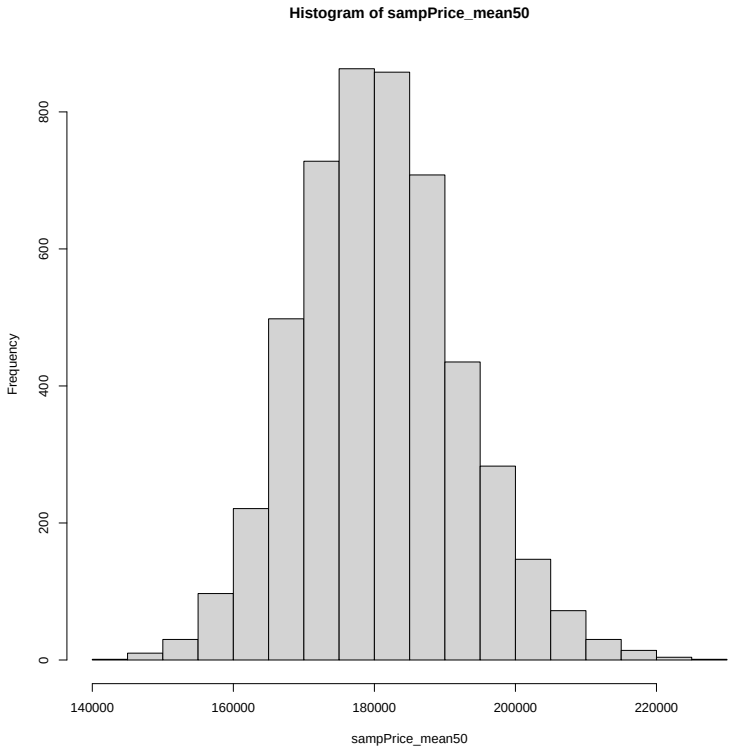
3 Opgaver

3.1 1

Vi tager en stikprøve med `@> sampPrice <- sample(price, 50)`, et punktestimat er:

```
@> mean(sampPrice)
[1] 176008.1
```

3.2 2



Figur 1: `Sample price mean`

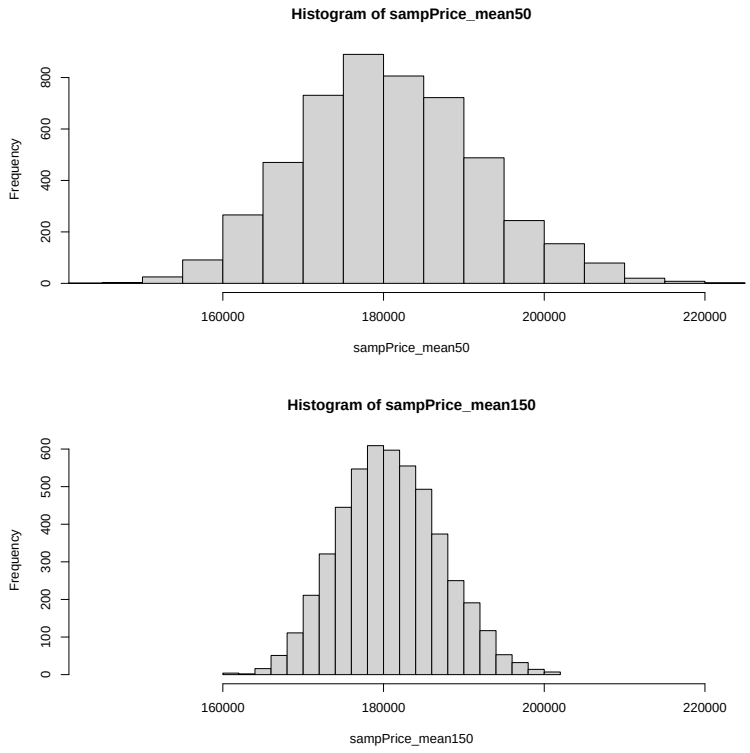
```
sampPrice_mean50 <- rep(NA, 5000)

for(i in 1:5000) {
  samp <- sample(price, 50)
  sampPrice_mean50[i] <- mean(samp)
}
```

Fordeling er normalfordelt, med en klokkeform. Ud fra histogrammet, kan det ses at middelværdien ligger omkring 18000$. Den reelle middelværdi for hele populationen er

```
@> mean(price)
[1] 180796.1
```

4 3

Vi laver en fordeling med stikprøvestørrelse på 150, for at kunne sammenligne:



Figur 2: Priser for et område med forskellige stikprøvestørrelser

Fordelingen med en stikprøvestørrelse på 150, har en pænere klokkeform end den med stikprøvestørrelse på 50. Spredningen for den med 150 prøver er mindre, og giver derfor prøve for prøve et bedre billede af data, end den med en stikprøvestørrelse på 50. Vi vil derfor gerne have en stikprøve af en hvis størrelse.