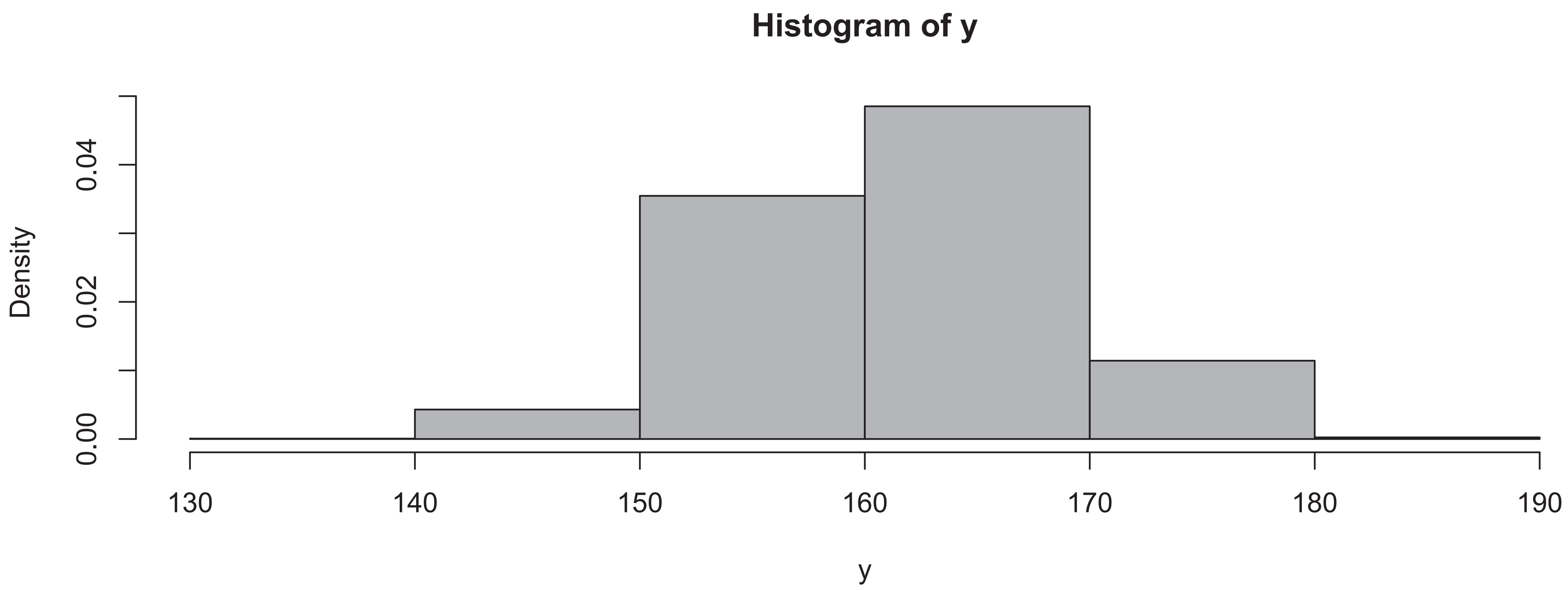
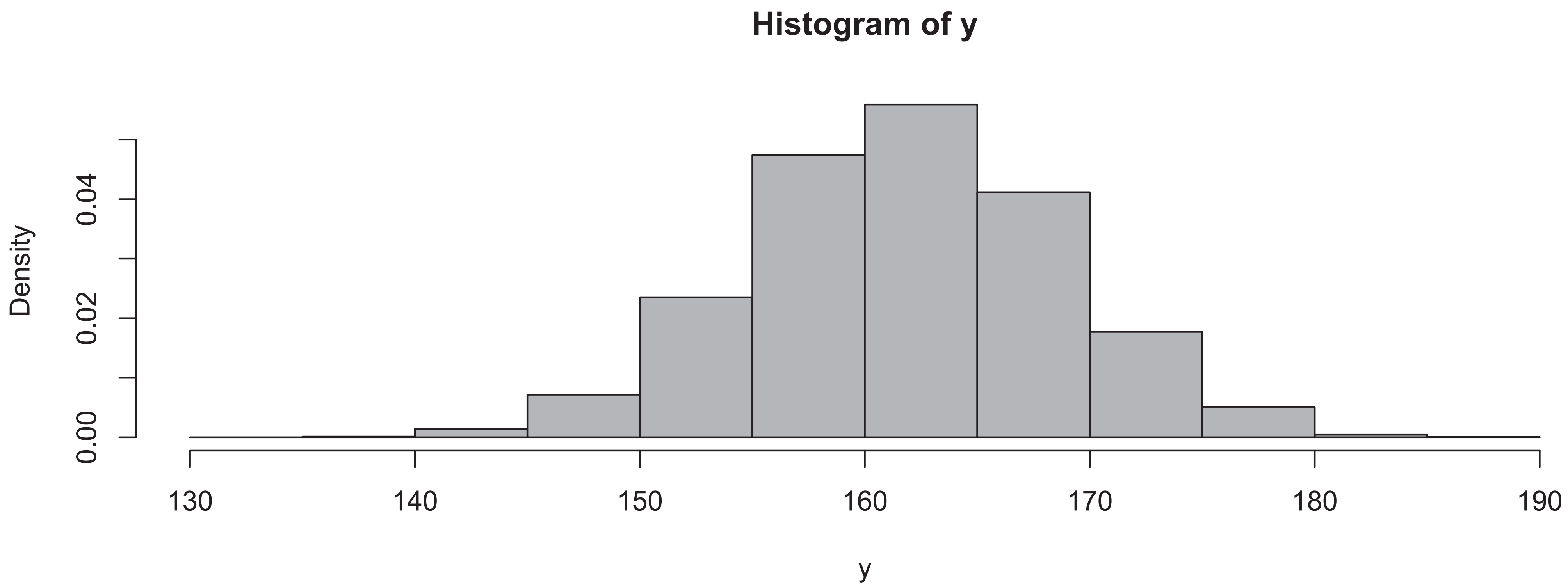
Hvorfor er normalfordelingen så normal?

Søren Højsgaard
Institut for Matematiske Fag, Aalborg Universitet

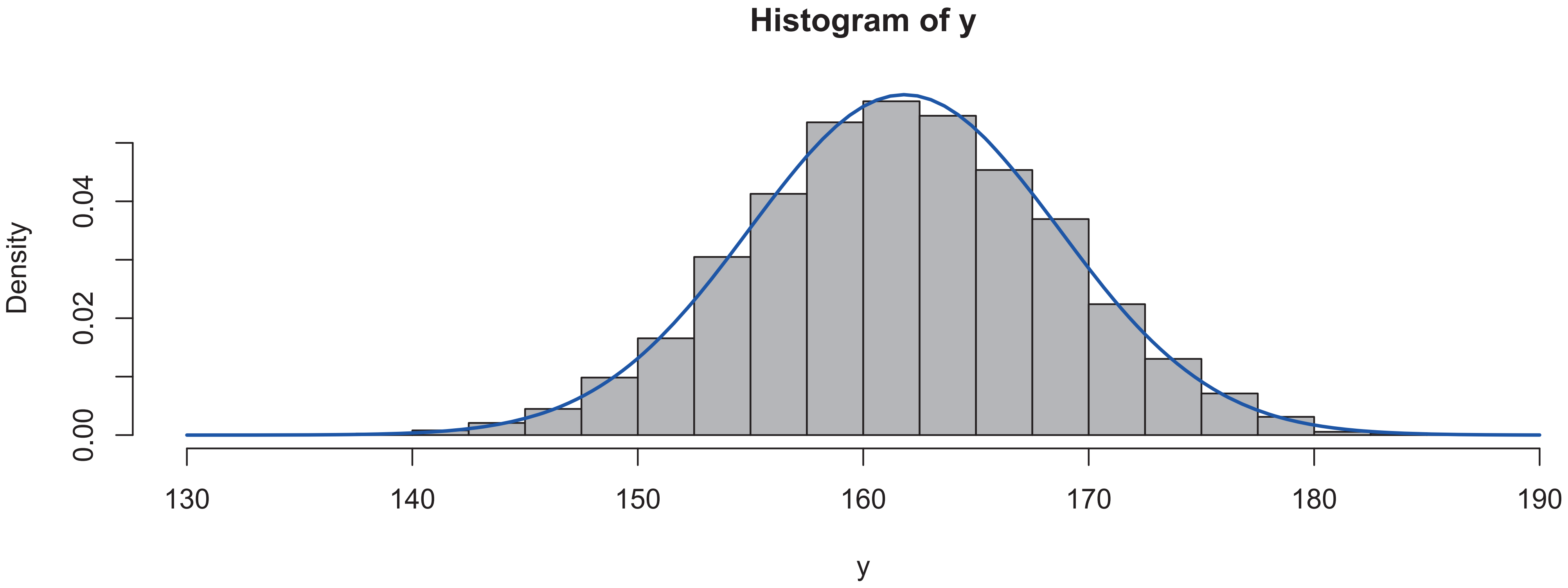
October 24, 2018

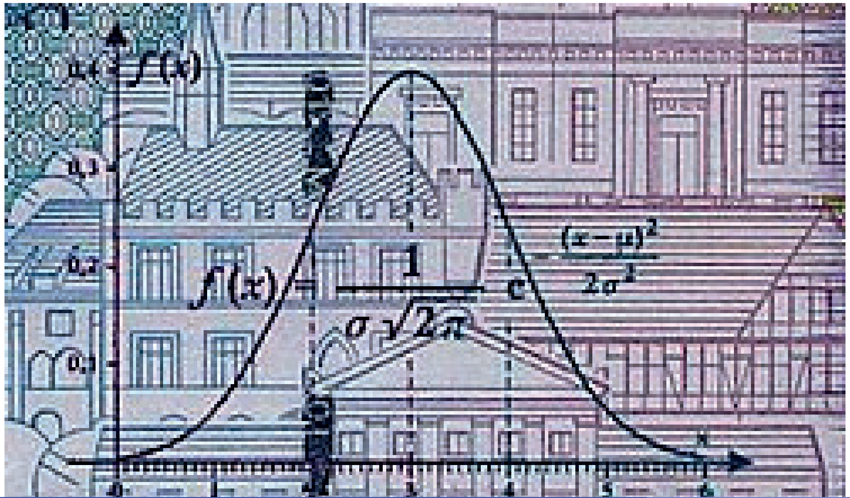
Højde af kvinder




Histogram of y
Density
0.00
0.02
0.04
130
140
150
160
170
180
190
y

Fortsætter man inddelingen i mindre grupper kan man forestille sig at histogrammet bliver en “glat klokkeform”.






GN4480100S8
DEUTSCHE BUNDESBANK
Banknote
10
ZEHN DEUTSCHE MARK
1777–1855 Carl Friedr. Gauß
Deutsche Bundesbank
Frankfurt am Main
1. September 1999
GN4480100S8

Figure 1:

Men hvorfor er normalfordelingen så normal?

Altså: Hvorfor kan så mange fænomener beskrives med en normalfordeling?

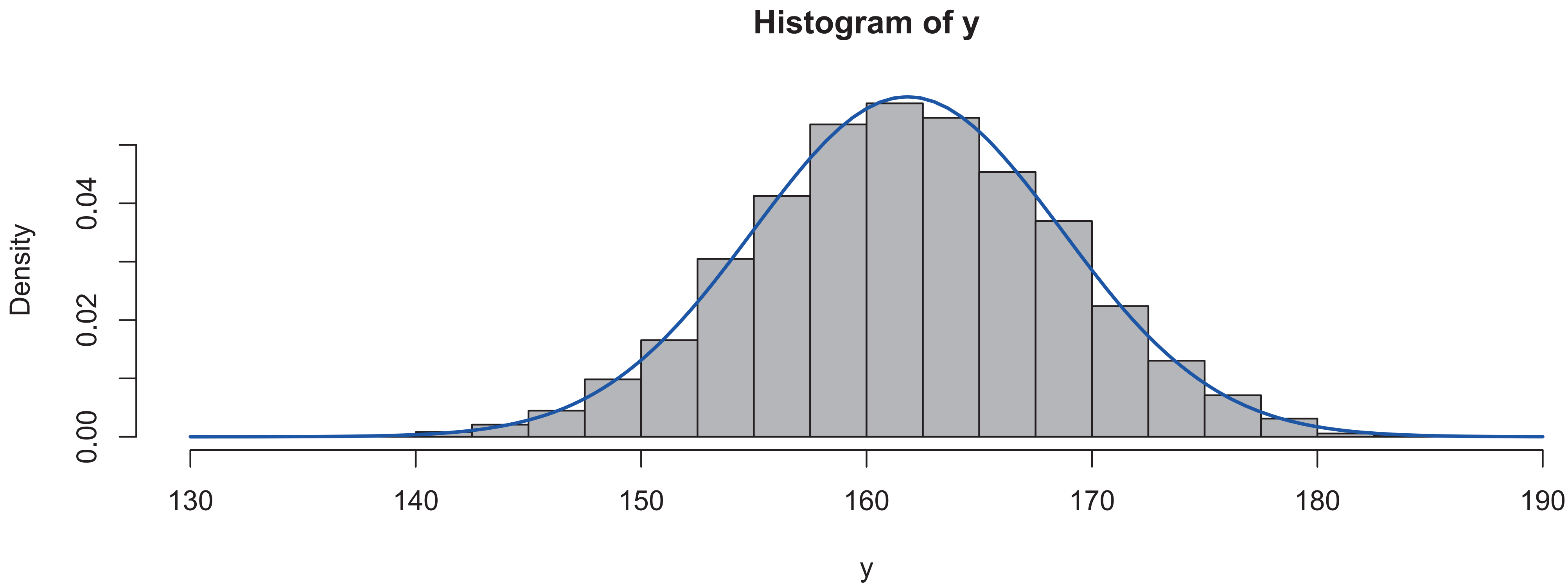
- Svaret ligger i den centrale grænseværdisætning - “the central limit theorem” eller CLT.
- Løst sagt: Summen af mange uafhængige bidrag er tilnærmelsesvist normalfordelt; og tilnærmelsen bliver bedre desto flere bidrag der er.

Hvad bestemmer en persons højde?

- Genetik; en bunke små bidrag
- Mad
- Levevilkår
- ...

I alt en mængde små (uafhængige) bidrag

Derfor bliver højden (med tilnærmelse) normalfordelt.



Den centrale grænseværdisætning

Der findes mange forskellige CLT'er; den simpleste er:

$X_1, X_2, \ldots, X_n$ uafhængige stokastiske variabler, hver med samme middelværdi $E(X_i) = \mu$ og samme varians $V(Y_i) = \sigma^2$.

Lad

$$S_n = \sum_i X_i, \quad Z_n = \frac{1}{n} S_n$$

Nemt at bevise, at

- $E(S_n) = n\mu$, $V(S_n) = n\sigma^2$.
- $E(Z_n) = \mu$, $V(Z_n) = \sigma^2/n$.

CLT fortæller mere: Når $n \to \infty$ så bliver fordelingen af S_n og Z_n **approximativt normal** (skrevet $\sim_A N(,)$)

$$S_n \sim_A N(n\mu, n\sigma^2), \quad Z_n \sim_A N(\mu, \sigma^2/n)$$

Hvordan simulerer man data fra en $N(\mu, \sigma^2)$-fordeling.

Obs: Hvis U er standard normalfordelt, $U \sim N(0, 1)$ så er $Y = \mu + \sigma U$ $N(\mu, \sigma^2)$-fordelt, så opgaven er at simulere fra $N(0, 1)$-fordeling.

Een måde: Lad Z være uniformt fordelt på $[-1/2; 1/2]$ (alle værdier i intervallet er lige sandsynlige; værdier udenfor intervallet forekommer ikke).

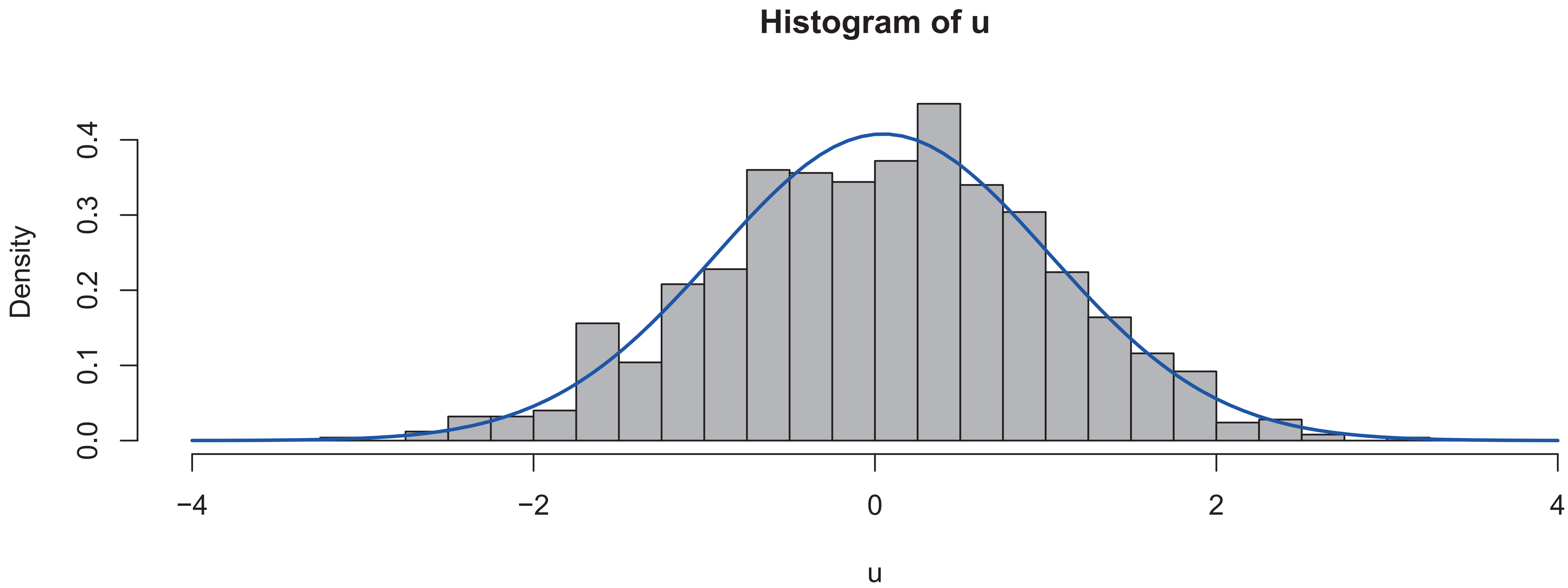
Kan klares med et lykkehjul :)

Obs: For en stokastisk variabel Z, der er uniformt fordelt $[-1/2; 1/2]$ er $E(Z) = 0$, $V(Z) = 1/12$.

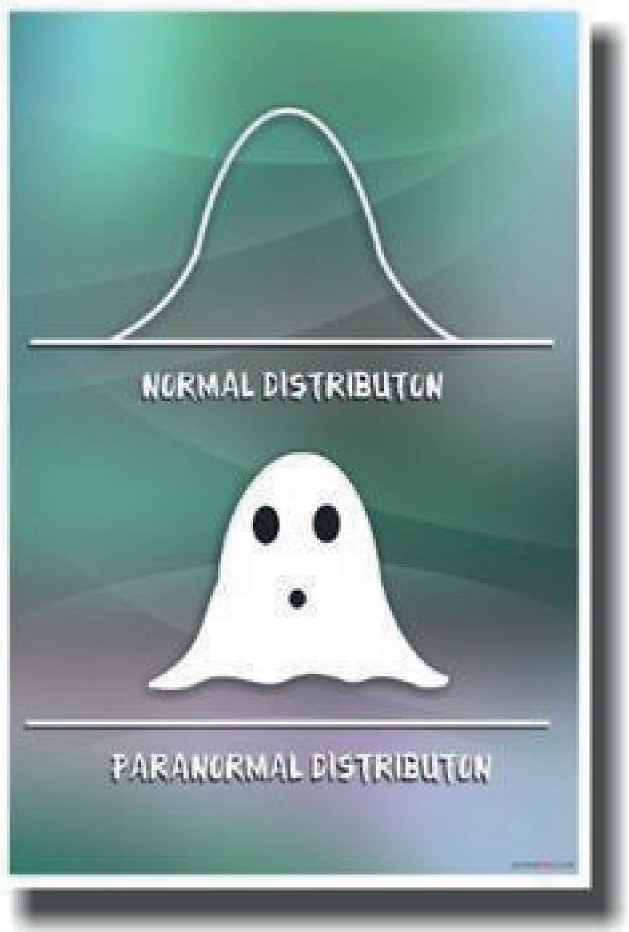
Lad $Z_1, \dots Z_{12}$ være uafhængige og uniforme på $[-1/2; 1/2]$.

Så er $U = \sum_{j=1}^{12} Z_j \sim_A N(0, 1)$

"Bevis:" Simuler U mange gange og tegn histogram; så bør man se klokkeformen:



Og måske er normalfordelingen slet ikke så normal alligevel. . .



Øvelse 1

Lad $Z_1, \ldots, Z_n$ være uafhængige of uniformt fordelte på intervallet $[-1/2; 1/2]$.

1. Summen af 2 uafhængige bidrag? Lad $U = \sum_{j=1}^{2} Z_j$ Simuler mange gange og tegn histogram. Ligner det en normalfordeling?
2. Summen af 4 uafhængige bidrag? Lad $U = \sum_{j=1}^{4} Z_j$ Simuler mange gange og tegn histogram. Ligner det en normalfordeling?
3. For en stokastisk variabel Z, der er uniformt fordelt $[-1/2; 1/2]$, vis at $E(Z) = 0$, $V(Z) = 1/12$.

Øvelse 2

Lad X være en binomialfordelt stokastisk variabel, $X \sim bin(N, \theta)$.

1. Find $E(X)$ og $V(X)$
2. Argumenter for, at X approximativt er normalfordelt for store N
3. Hvad er den approximative fordeling af $\frac{X}{N}$?
4. Hvad er den approximative fordeling af $\frac{\frac{X}{N} - \theta}{\sqrt{\theta(1-\theta)/N}}$?