Oefeningen hoofdstuk 4 - steekproefonderzoek

Tijs Martens

13 maart 2019

Voorbeelden blz 55

oefeningen gebaseerd op grafiek 4.3. blz 53

voorbeeld 4.3.

```
pnorm(4, 5, 1.5)
```

```
## [1] 0.2524925
```

voorbeeld 4.4.

```
pnorm(7, 5, 1.5)
```

```
## [1] 0.9087888
```

voorbeeld 4.5.

```
pnorm(3, 5, 1.5)
```

```
## [1] 0.09121122
```

voorbeeld 4.5.

```
ondergrens <- pnorm (2, 5, 1.5)
bovengrens <-pnorm (6.5, 5, 1.5)

bovengrens - ondergrens
```

```
## [1] 0.8185946
```

oefening 4.1.

Een onderzoeker wil zo correct mogelijk de consumptiegewoontes van de inwoners van 18 jaar en ouder in een bepaalde gemeente, met 3 woonkernen, onderzoeken. Hij onderscheidt 4 leeftijdsgroepen zodat hij uiteindelijk aan 12 deelgroepen komt. Hij vraagt de procentuele samenstelling van de bevolking op in de gemeente en berekent daaruit hoeveel bevragingen hij per deelgroep moet uitvoeren. Dit noemen we een quotasteekproef.

a. voor en nadelen

voordelen:

- representatief
- per segment informatie beschikbaar

nadelen:

- kost
- tijd

b. welke fouten kunnne er optreden

toevallige steekproeffouten

- door toeval de verkeerde mensen selecteren die voor een niet representatief gemiddelde zorgen

systematische steekproeffouten

- te veel mensen van een groep bevragen en geen to niet genoeg van een andere groep

toevallige niet steekproeffouten

- de vraag wordt verkeerd geïnterpreteerd
- het antwoord wordt verkeerd geformuleer

systematische niet steekproeffouten

- als de bevraagde voordeel halen uit de antwoorden
- verkeerde zaken noteren
- verkeerde categoriën bepaald (categorie = strata)

c welke andere parameters zouden kunnen gebrukt worden bij het opslitsen in deelgroepen?

- financiele situatie
- kinderen
- job
- burgerlijke stand
- leeftijd
- ...

oefening 4.2.

(zelf gemaakt, niet de oplossing)

a.

dit is een niet-aselecte steekproef. Er zijn objecten die geen kans hebben om in de steekproef voor te komen.

- het is 2019, mannen kunnne ook de was doen
- er zijn jongere en oudere mensen die de was doen

deze steekproeffout is een "systematische steekproeffout"

b.

groot, we kunnne geen conclussies trekken als er geen aselecte steekproef is

oefening 4.3.

(zelf gemaakt, niet de oplossing)

a.

Als er niet in elke vestiging even veel werknemers werken is de kans dat een persoon gekozen wordt uit een grote vestiging kleiner. Er zou procentueel gekeken moeten worden naar de omvang van een vestiging en op basis daarvan bepalen hoeveel werkenemers er per vestiging gekozen worden.

b.

Als alle vestigingen even groot zijn.

oefening 4.4.

(zelf gemaakt, niet de oplossing)

a.

- systematische steekproeffout: niet elke student heeft kans om in het experiment opgenomen te worden
- De studenten van een bepaalde richting hebben niet dezelfde cultuur/ mindset als de volledige HoGent. Je kan een richting dus niet veralgemenen.
- Studenten die niet aanwezig zijn, zijn ook objecten die deel uitmaken van de populatie. Deze hebben geen kans om gekozen te worden.

b.

Studenten kunnen geintimideerd worden door de docent en op die manier kunnne de antwoorden be?nvloed worden.

c.

Studenten met een slecht resultaat zullen en slecht gevoel hebben tegenover de faculteit. Dit zal invloed hebben op het antwoord.

oen te berekenen voor de standaardnormale verdeling .

Z is normaalverdeeld: $N(0; 1)$.

oefening 4.5.

a.

$P(Z < 1.33)$

```
kans <- pnorm(1.33)
kans
```

```
## [1] 0.9082409
```

b.

$P(Z > 1.33)$

```
kans <- 1 - pnorm(1.33)
kans
```

```
## [1] 0.09175914
```

c./ d.

- wegens symetrie is $P(Z < 1.33) = P(Z > 1.33)$
- wegens complementregel is $P(Z > -1.33) = 1 - P(Z < -1.33)$

e.

```
kans <- pnorm(0.45)
kans
```

```
## [1] 0.6736448
```

f.

```
kans <- pnorm(1.05)
kans
```

```
## [1] 0.8531409
```

g.

```
kans <-pnorm(0.65)
kans
```

```
## [1] 0.7421539
```

h.

$P(-0.45 < Z < 1.20)$

$$P(-0.45 < Z < 1.20) = 1 - P(Z < -0.45) - P(Z < 1.20)$$

```
kans <- 1 - pnorm(-0.45) - pnorm(-1.20)
kans
```

```
## [1] 0.5585751
```

$P(-1.35 < Z < -0.10)$

```
kans1 <- pnorm(-1.35)
kans2 <- pnorm(-0.10)
kans1
```

```
## [1] 0.08850799
```

```
kans2
```

```
## [1] 0.4601722
```

```
result <- kans2-kans1
result
```

```
## [1] 0.3716642
```

oefening 4.6.

(zelf gemaakt, niet de oplossing)

opgave

Bepaal de dichtheid en de cumulatieve waarschijnlijkheidscurve voor een normale verdeling met een gemiddelde μ van 2,5 en σ= 1,5. Bepaal de oppervlakte voor het gebied onder de dichtheidscurve tussen x = 0.5 en x = 4. Controleer uw antwoord door de berekening te doen.

oplossing

algmeeen

```
mu <- 2.5
sigma <- 1.5

minwaarde = 0.5
maxwaarde = 4
```

bepalen dichtheid

```
ondergrens <- pnorm(q = minwaarde, mean = mu, sd = sigma)
bovengrens <- pnorm(q = maxwaarde, mean = mu, sd = sigma)

bovengrens - ondergrens
```

```
## [1] 0.7501335
```

maken cumulatieve waarschijnlijkheidscurve

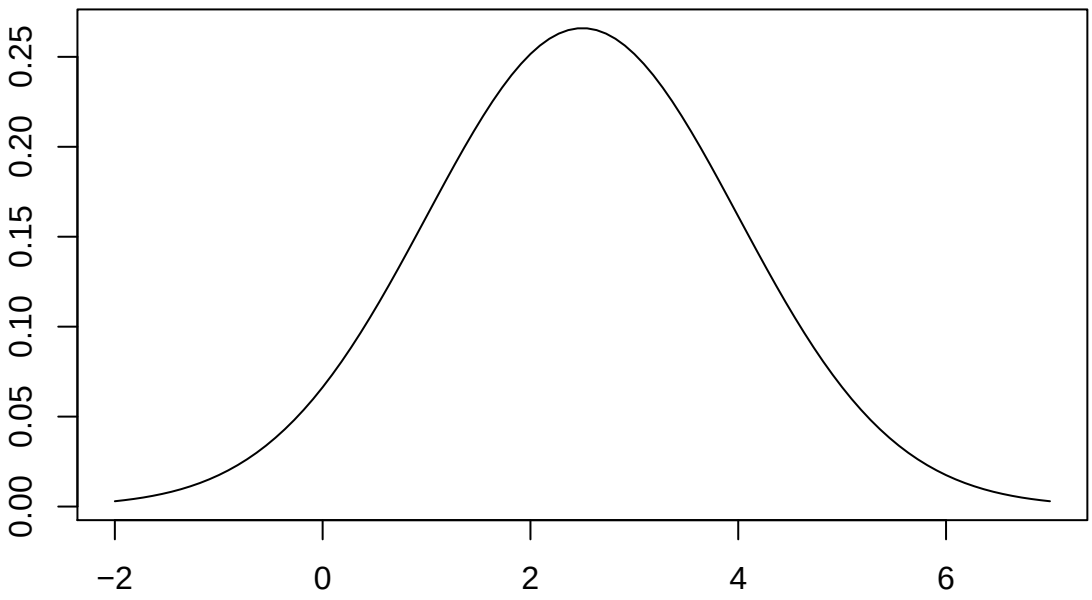
```
x <- seq(-2, 7, length = 100)
x
```

```
##   [1] -2.00000000 -1.90909091 -1.81818182 -1.72727273 -1.63636364
##   [6] -1.54545455 -1.45454545 -1.36363636 -1.27272727 -1.18181818
##  [11] -1.09090909 -1.00000000 -0.90909091 -0.81818182 -0.72727273
##  [16] -0.63636364 -0.54545455 -0.45454545 -0.36363636 -0.27272727
##  [21] -0.18181818 -0.09090909  0.00000000  0.09090909  0.18181818
##  [26]  0.27272727  0.36363636  0.45454545  0.54545455  0.63636364
##  [31]  0.72727273  0.81818182  0.90909091  1.00000000  1.09090909
##  [36]  1.18181818  1.27272727  1.36363636  1.45454545  1.54545455
##  [41]  1.63636364  1.72727273  1.81818182  1.90909091  2.00000000
##  [46]  2.09090909  2.18181818  2.27272727  2.36363636  2.45454545
##  [51]  2.54545455  2.63636364  2.72727273  2.81818182  2.90909091
##  [56]  3.00000000  3.09090909  3.18181818  3.27272727  3.36363636
##  [61]  3.45454545  3.54545455  3.63636364  3.72727273  3.81818182
##  [66]  3.90909091  4.00000000  4.09090909  4.18181818  4.27272727
##  [71]  4.36363636  4.45454545  4.54545455  4.63636364  4.72727273
##  [76]  4.81818182  4.90909091  5.00000000  5.09090909  5.18181818
##  [81]  5.27272727  5.36363636  5.45454545  5.54545455  5.63636364
##  [86]  5.72727273  5.81818182  5.90909091  6.00000000  6.09090909
##  [91]  6.18181818  6.27272727  6.36363636  6.45454545  6.54545455
##  [96]  6.63636364  6.72727273  6.81818182  6.90909091  7.00000000
```

```
y <- qnorm(x)
```

```
## Warning in qnorm(x): NaNs produced
```

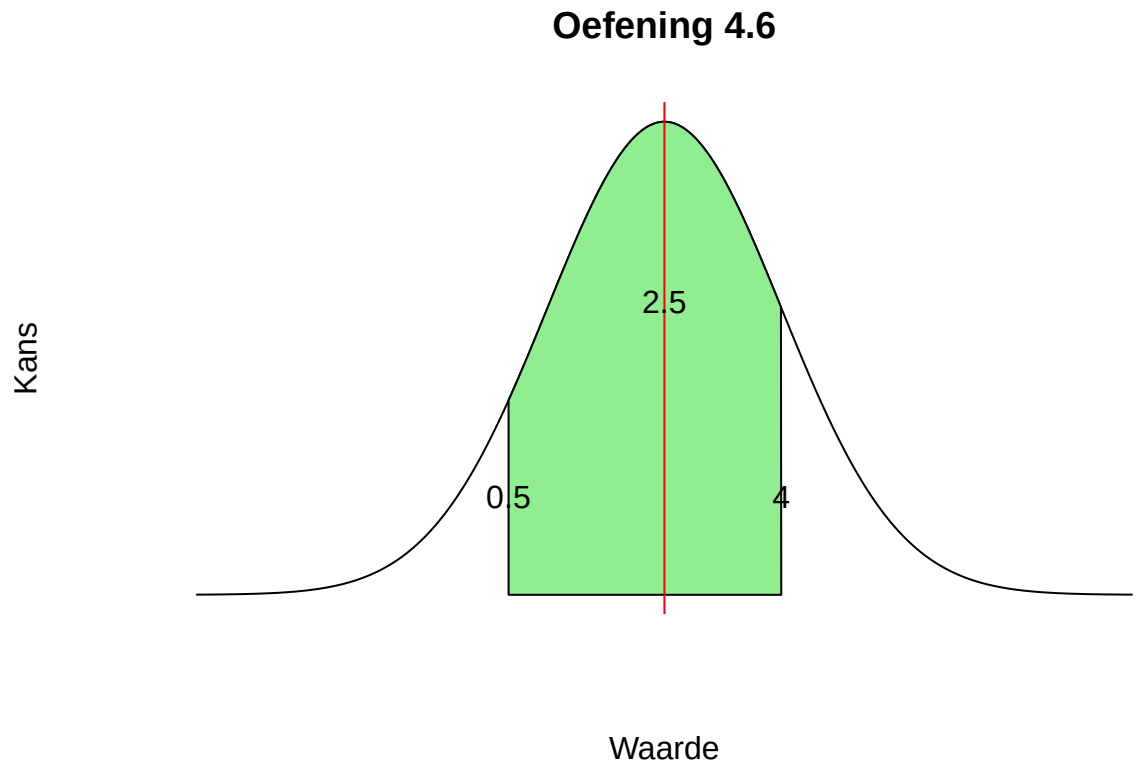
```
dist <- dnorm(x, mu, sigma)

plot(x, dist, type="l", xlab = '', ylab = '')
```



bepalen oppervlake voor het gebied tussen x = 0.5 en x = 4

```
population_mean <- 2.5
population_sd <- 1.5

sd_to_fill <- 1

lower_bound <- 0.5
upper_bound <- 4

x <- seq(-4, 4, length = 1000) * population_sd + population_mean
y <- dnorm(x, population_mean, population_sd)

plot(x, y, type="n", xlab = "Waarde", ylab = "Kans", main = "Oefening 4.6", axes = FALSE)
lines(x, y)

bounds_filter <- x >= lower_bound & x <= upper_bound
x_within_bounds <- x[bounds_filter]
y_within_bounds <- y[bounds_filter]
x_polygon <- c(lower_bound, x_within_bounds, upper_bound)
y_polygon <- c(0, y_within_bounds, 0)

polygon(x_polygon, y_polygon, col = "lightgreen")

abline(v=population_mean, col='red')

text(population_mean, mean(dist)*1.5, population_mean)
```

```
text(lower_bound,mean(dist)*0.5,signif(lower_bound, digits=4))
text(upper_bound,mean(dist)*0.5,signif(upper_bound, digits=4))
```



controlleer je antwoord door de berekening te doen

```
bovengrens - ondergrens
```

```
## [1] 0.7501335
```

oefening 4.7.

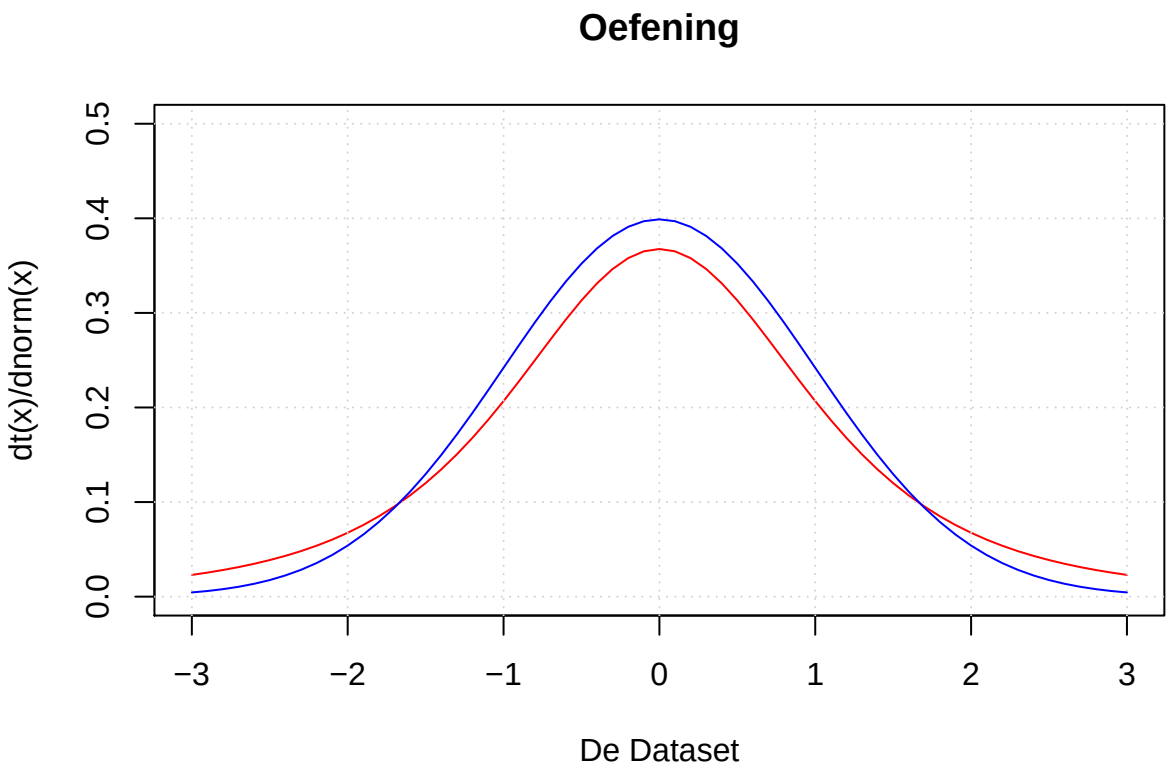
opgave

Bepaal de dichtheid en de cumulatieve waarschijnlijkheidscurve voor een t-verdeling met d f = 3. Teken ook een normale verdeling met een $\mu = 0$ en $\sigma = 1$.

oplossing

```
dataSet <- seq(-3, 3, by=0.1)
y <- dt(dataSet, df=3)

plot(dataSet, y, type="l", col="red", ylim= c(0, 0.5), xlab = "De Dataset", ylab = "dt(x)/dnorm(x)", ma

grid()
```

```
yNorm <- dnorm(dataSet)
lines(x = dataSet, y = yNorm, type='l', col="blue")
```



tweede poging oplossing

dichtheid

cummulatieve waarschijnlijkheid

normale verdeling met mu = 0 en sigma = 1

oefening 4.8.

(zelf gemaakt, niet de oplossing)

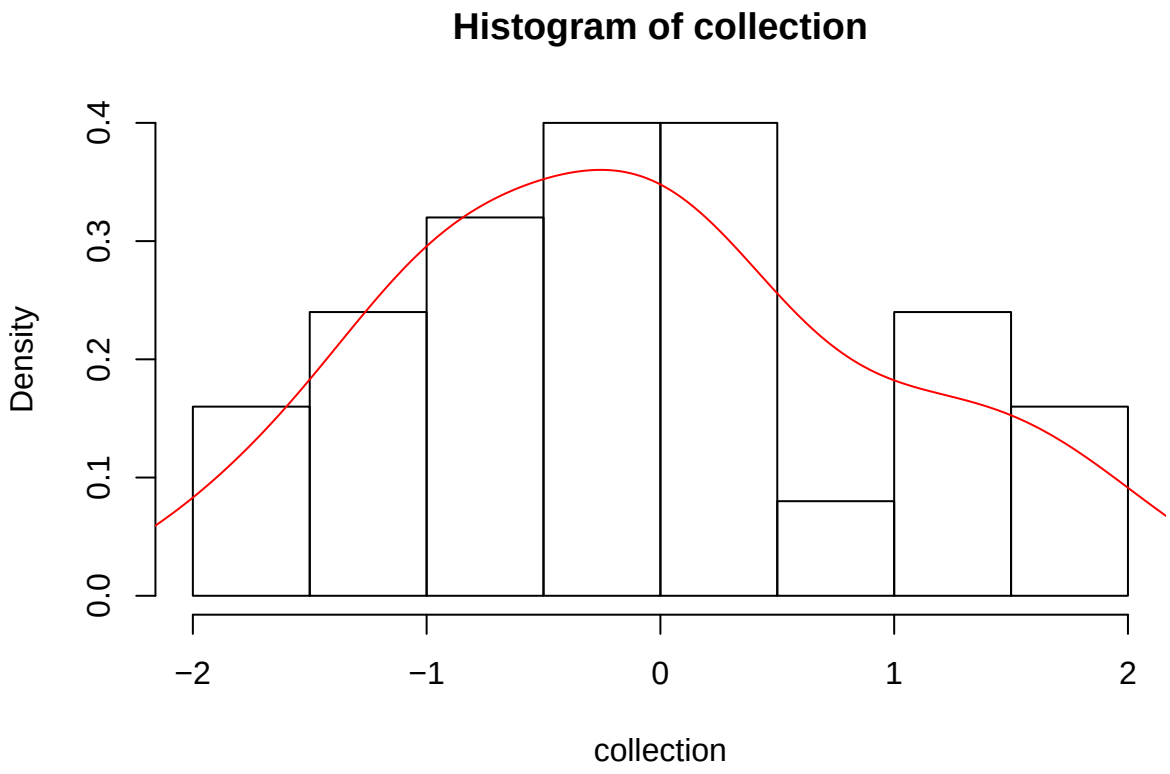
opgave

Gebruik de functie rnorm() een willekeurige steekproef van 25 waarden uit een normale verdeling te tekenen met een gemiddelde van 0 en een standaardafwijking gelijk aan 66 Hoofdstuk 4. Steekproefonderzoek 1,0. Gebruik een histogram, met probability = TRUE. Maak een overlay over het histogram met: (a) de theoretische dichtheidscurve voor een normale verdeling met gemiddelde 0 en standaardafwijking gelijk aan 1,0; (b) een “geschatte” dichtheidscurve op basis van het gemeten steekproefgemiddelde en -standaardafwijking. Herhaal dit voor een steekproef van 100 en 500 waarden.

oplossing

voor 25:

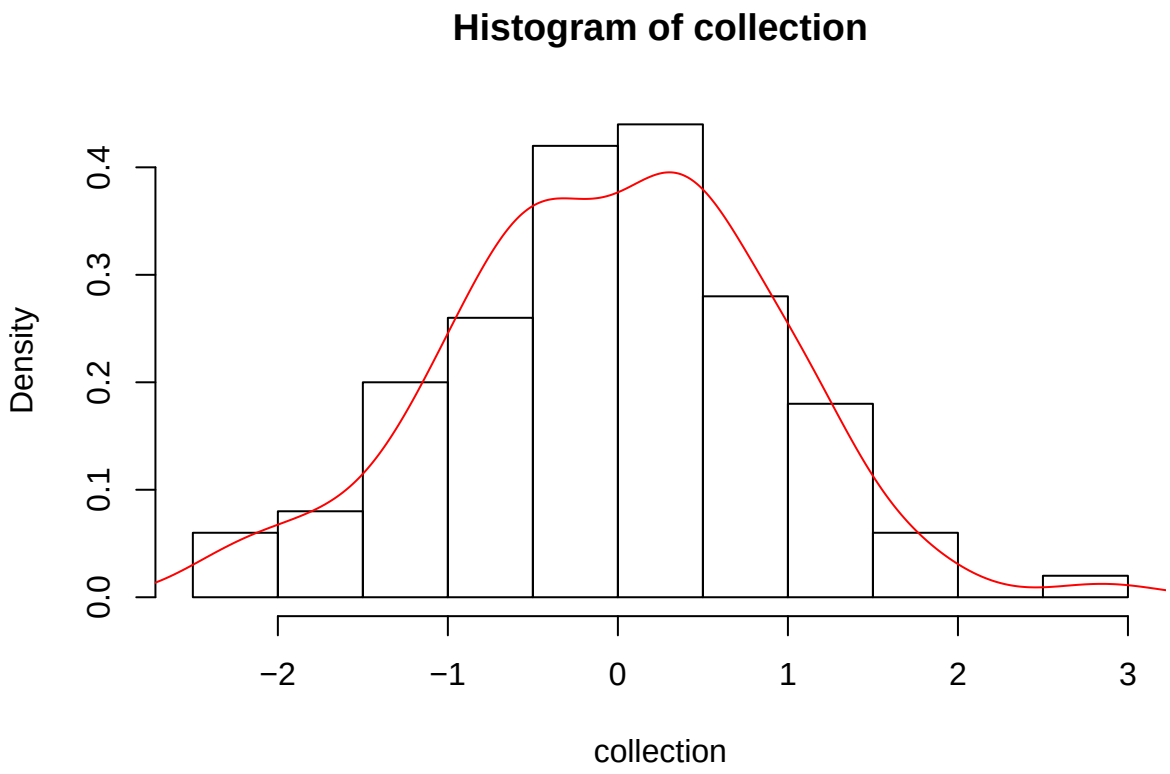
```r
collection <- rnorm(25, mean = 0, sd = 1.0)



hist(x = collection, probability = TRUE)
#proubability =TRUE zorgt voor de rode lijn
lines(density(collection), col='red')
```




```r
## geen idee hoe ik theoretische gaus curve moet toevoegen
```

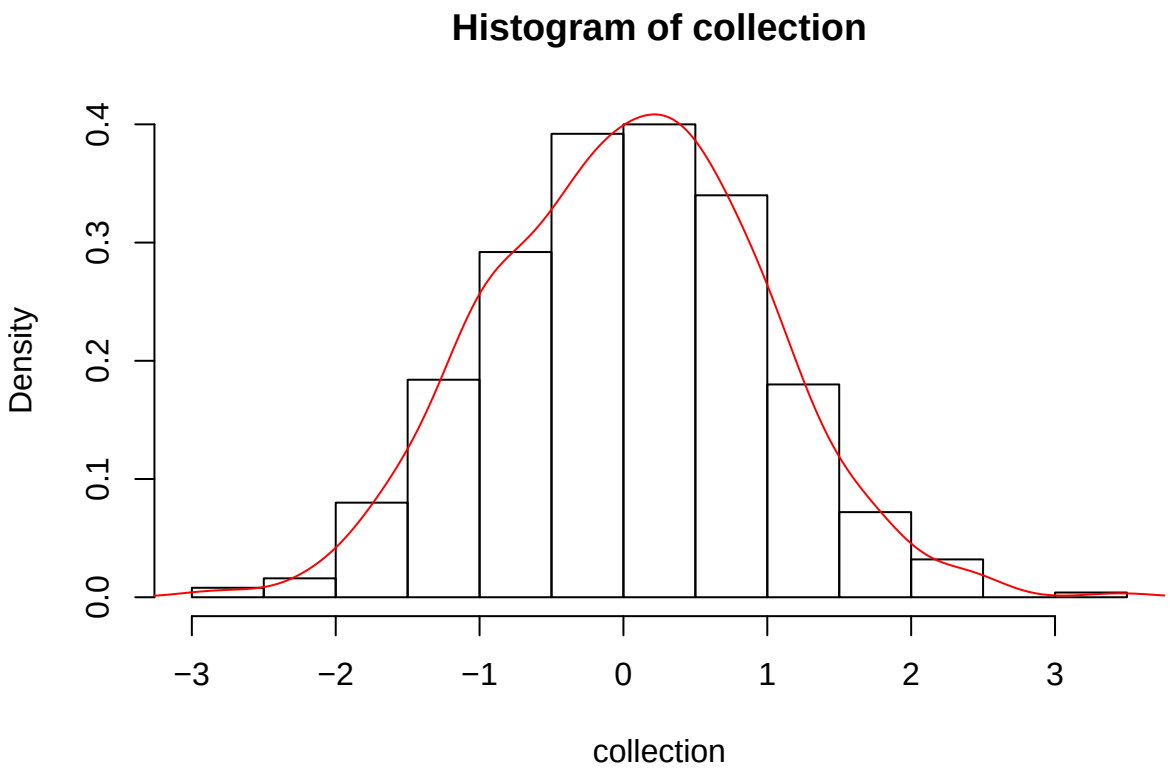
voor 100:

```r
collection <- rnorm(100, mean = 0, sd = 1.0)



hist(x = collection, probability = TRUE)
lines(density(collection), col='red')
```



voor 500:

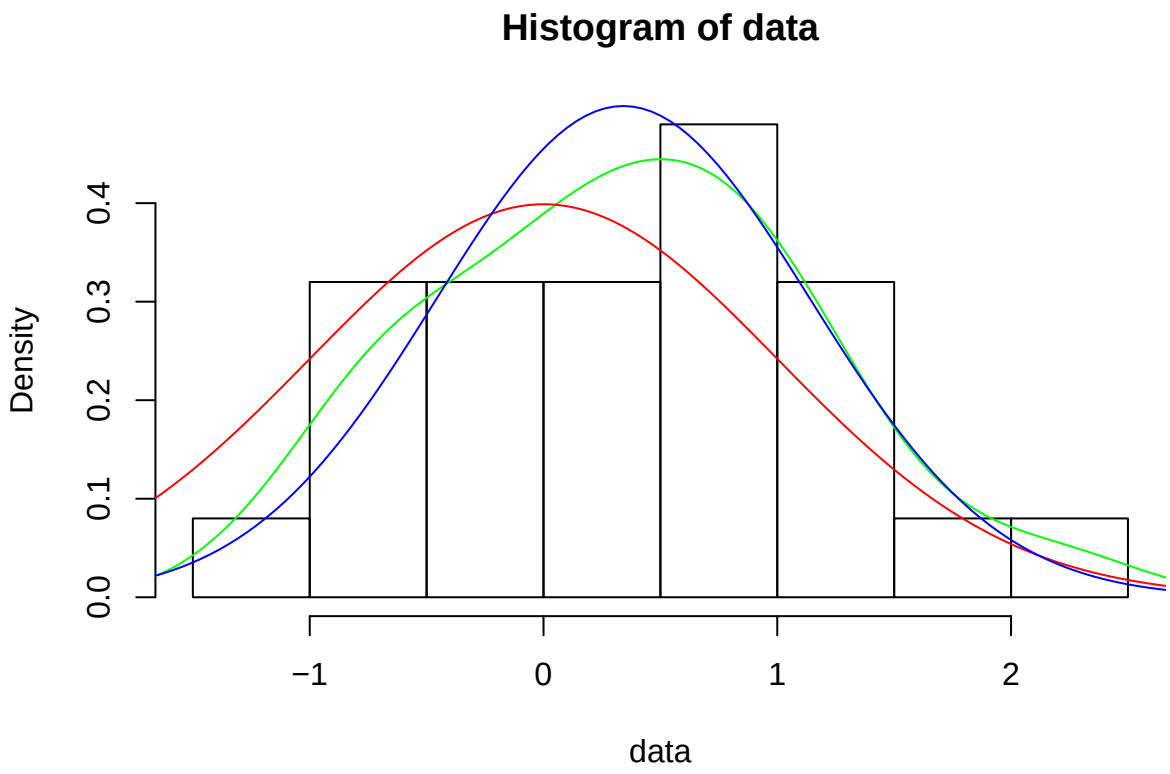
```
collection <- rnorm(500, mean = 0, sd = 1.0)



hist(x = collection, probability = TRUE)
lines(density(collection), col='red')
```



tweede poging oplossing

voor 25:

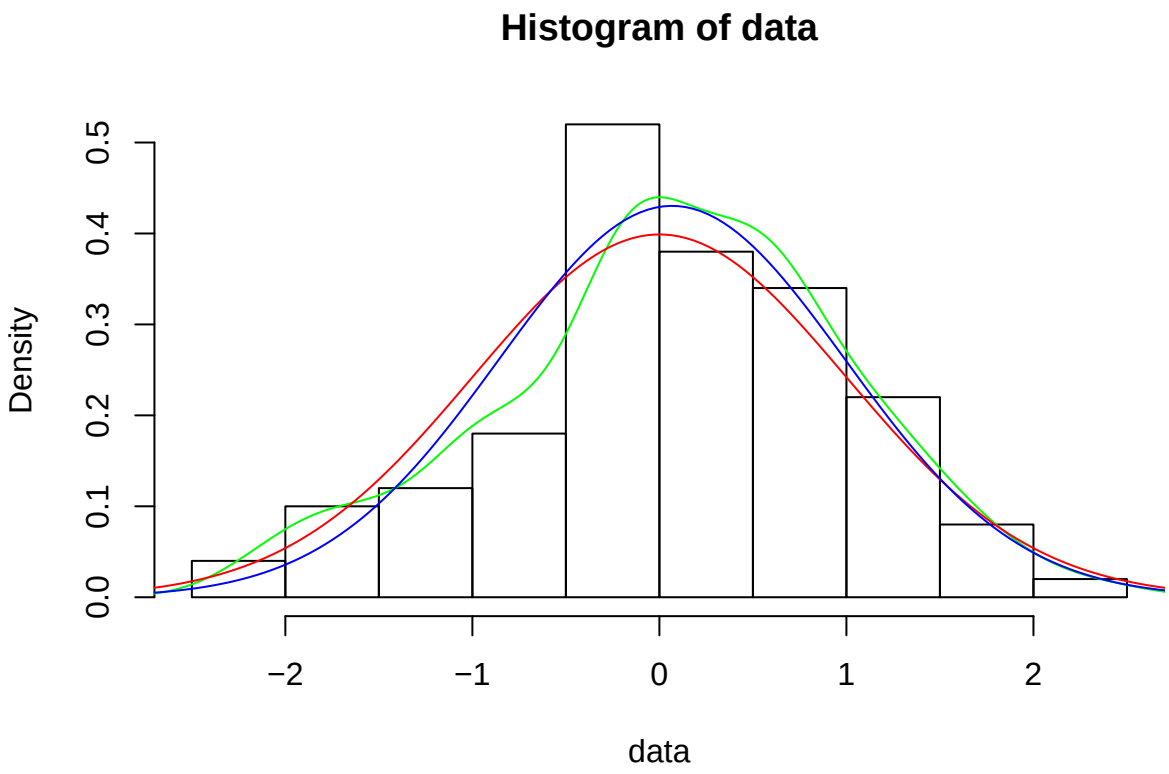
```
aantal <- 25
gem <- 0
sig <- 1

data <- rnorm(aantal, mean = gem, sd = sig)

steekproefgem <- mean(data)
steekproefsd <- sd(data)


x <- seq(from = gem - 4 * sig,
         to = gem + 4 * sig,
         length.out = 200)

y_theoretisch <-  dnorm(x, gem, sig)
y_steekproef <- dnorm(x, steekproefgem, steekproefsd)

hist(x = data, probability = TRUE)

lines(density(data), col='green')
lines(x, y_theoretisch, col='red')
lines(x, y_steekproef, col="blue")
```



groen = dichtheid rood = curve van de volledige populatie blauw = curve van de steekproef

voor 100:

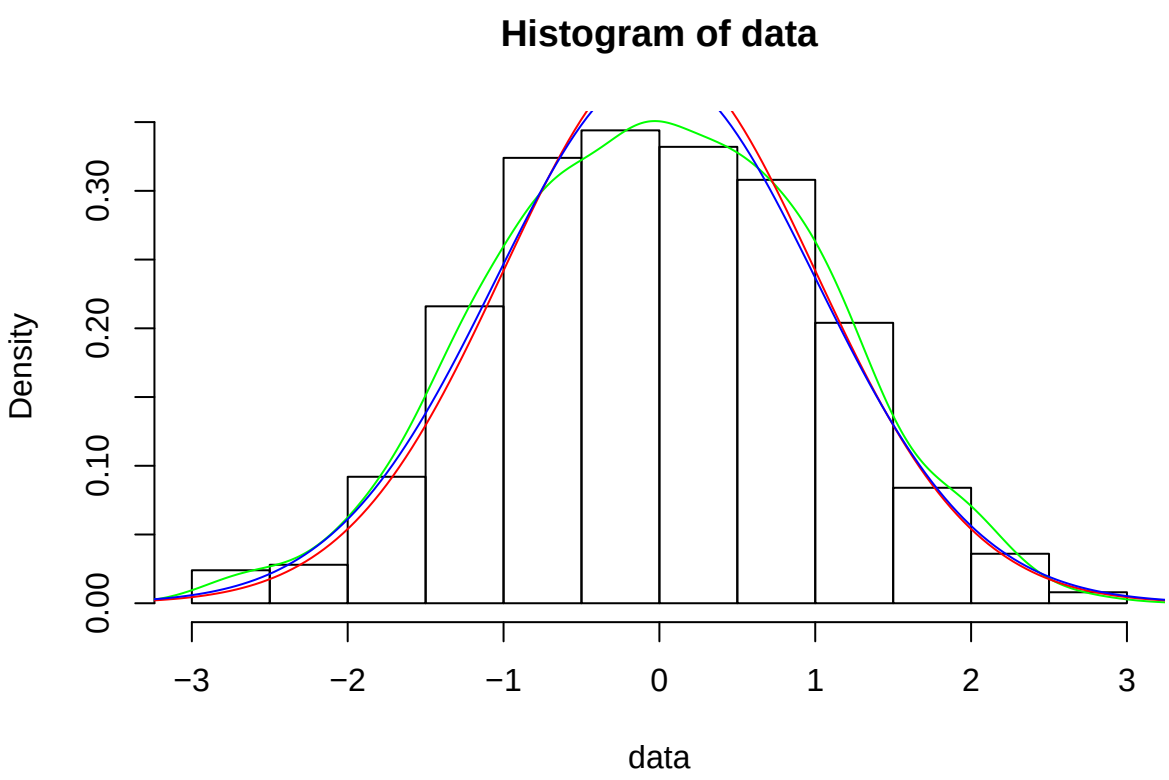
```
aantal <- 100
gem <- 0
sig <- 1

data <- rnorm(aantal, mean = gem, sd = sig)

steekproefgem <- mean(data)
steekproefsd <- sd(data)


x <- seq(from = gem - 4 * sig,
         to = gem + 4 * sig,
         length.out = 200)

y_theoretisch <-  dnorm(x, gem, sig)
y_steekproef <- dnorm(x, steekproefgem, steekproefsd)

hist(x = data, probability = TRUE)

lines(density(data), col='green')
lines(x, y_theoretisch, col='red')
lines(x, y_steekproef, col="blue")
```



groen = dichtheid rood = curve van de volledige populatie blauw = curve van de steekproef

voor 25:

```
aantal <- 500
gem <- 0
sig <- 1

data <- rnorm(aantal, mean = gem, sd = sig)

steekproefgem <- mean(data)
steekproefsd <- sd(data)


x <- seq(from = gem - 4 * sig,
         to = gem + 4 * sig,
         length.out = 200)

y_theoretisch <-  dnorm(x, gem, sig)
y_steekproef <- dnorm(x, steekproefgem, steekproefsd)

hist(x = data, probability = TRUE)

lines(density(data), col='green')
lines(x, y_theoretisch, col='red')
lines(x, y_steekproef, col="blue")
```



groen = dichtheid rood = curve van de volledige populatie blauw = curve van de steekproef

oefening 4.9.

opgave

In de Hogeschool zijn er twee klassen voor het vak onderzoekstechnieken. De studenten werden willekeurig over de klassen verdeeld, zodat we mogen veronderstellen dat de ene klas niet slimmer is dan de andere. In de A-klas geeft mevr. X les, in de B-klas geeft mr. Y les. X is nogal streng en op het einde van het schooljaar behaalt haar klas een gemiddelde van 54 op 100 met een standaardafwijking van 11. Y is iets losser en stimuleert de leerlingen al gauw met een puntje meer. Op het einde van het schooljaar behaalt zijn klas een gemiddelde van 62 op 100 en een standaardafwijking van 7. Wouter zit in de A-klas en heeft 63/ 100 voor wiskunde. Stijn zit in de B-klas en behaalt 67/100 . Wie heeft volgens jou het beste gescoord binnen de eigen klas?

oplossing

Wouter heeft een betere nominale (z-score) dan Stijn. Wouter heeft dus beter gescoord

formule z-score: $z = (x - \mu)/\sigma$

- x = gemeten waarden
- μ = gemiddelde
- sigma = standaardaafwijking

```
z_wouter <- (63-54)/11
z_stijn <- (67-62)/7
```

```
z_wouter
```

```
## [1] 0.8181818
```

```
z_stijn
```

```
## [1] 0.7142857
```

Wouter heeft een z-score van 0.82 Stijn heeft een z-score van 0.71

een z score van 1 wil zeggen dat de geobserveerde waarde 1 standaardafwijking van het gemiddeld ligt

Aangezien de z-score van Wouter hoger is dan die van Stijn veronderstellen we dat Wouter relatief beter scoort.

Dit weten we nu nog niet zeker. We weten enkel dat Wouter verder van het gemiddelde verwijderd is.

Aangezien de geobserveerde waarde groter is dan het gemiddelde van de klas kunnen we afleiden dat beide scores boven het gemmiddelde ligt.

Wouter heeft dus relatief het beste gescoord.

oefening 4.10.

opgave

Een gezondheidsonderzoek tussen 1988 en 1994 gaf aan dat de gemiddelde cholesterolwaarde bij vrouwen tussen 20 en 29 jaar 183 mg/dl bedroeg, met een standaardafwijking gelijk aan 36. We nemen nu een aselecte steekproef van 81 vrouwen. Los volgende vragen op:

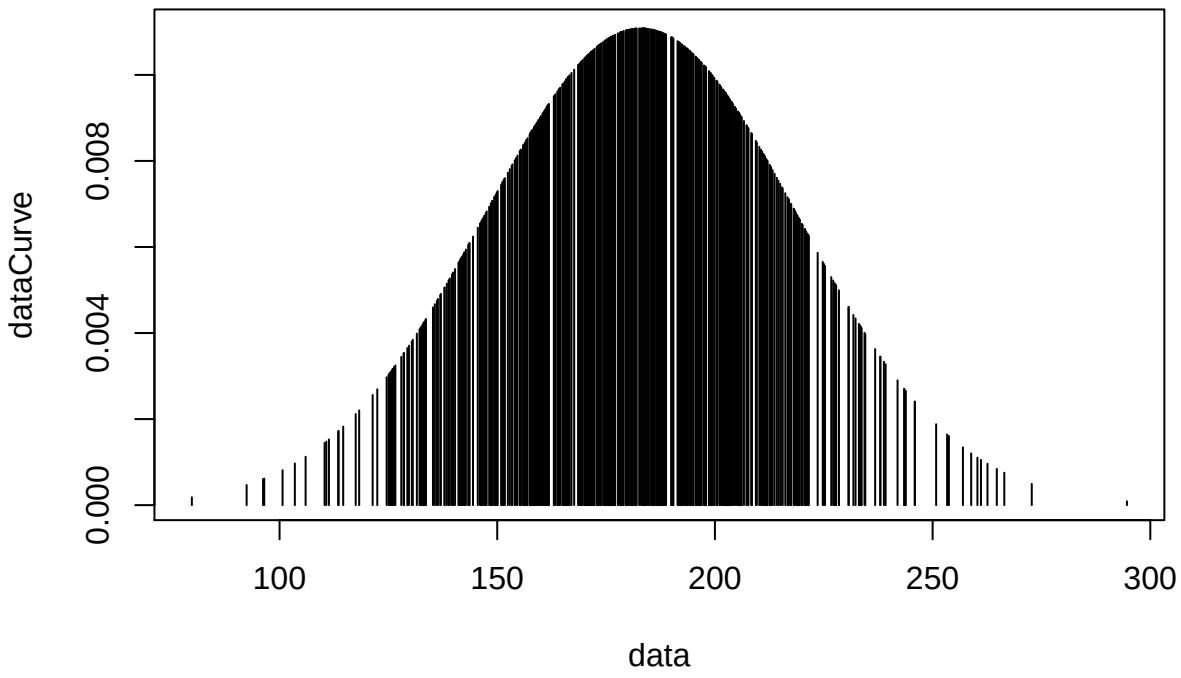
1. Schets de kansdichtheidsfunctie voor de populatie en de kansverdeling van het steekproefgemiddelde x.
2. Bepaald de kans dat x kleiner is dan 185.
3. Bepaal de kans dat x tussen 175 en 185 ligt.
4. Bepaal de kans dat x groter is dan 190.

oplossing

a.

zie cursusblad

```
gem <- 183
sd <- 36
n <- 81

data <- rnorm(500,gem, sd)

steekproefGemiddelde <- mean(data)
steekproefSd <- sd

dataCurve <- dnorm(data, gem, sd)
## bestaat uit 500 punten

plot(data, dataCurve, type = 'h')
```



b.

$p(\overline{x} \leq 185)$

```
mu <- 183
sd <- 36
n <- 81

prob <- pnorm(185, mu, sd/sqrt(n))
prob
```

```
## [1] 0.6914625
```

c.

$P(175 \leq \overline{x} \leq 185)$

```
prob <- pnorm(185, mu, sd/sqrt(n)) - pnorm(175,mu, sd/sqrt(n))
prob
```

```
## [1] 0.6687123
```

d.

$P(\overline{x} \geq 190)$

```
prob <- 1-(pnorm(190, mu, sd/sqrt(n)))
prob
```

```
## [1] 0.04005916
```

oefening 4.11.

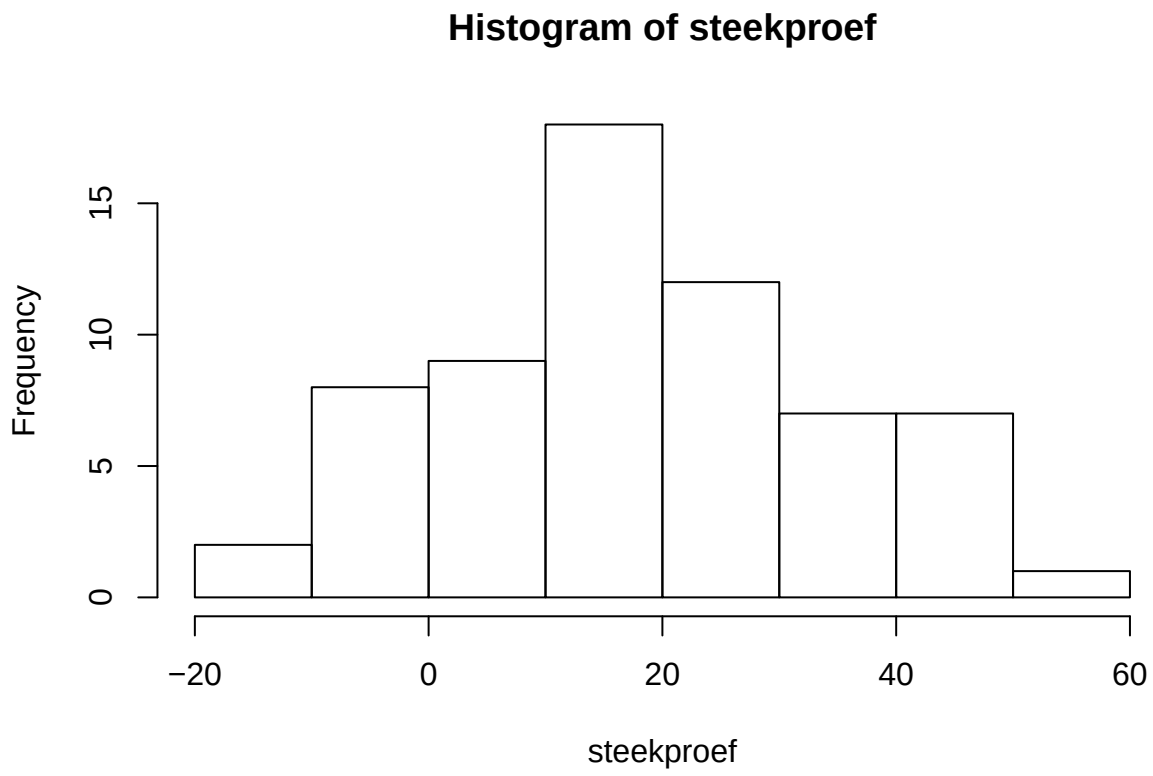
(zelf gemaakt, niet de oplossing)

opgave

zie blz 66

oplossing

```
steekproef <- rnorm(64, mean = 20, sd = 16)

hist(steekproef)
```



A.

```
afwijking <- sd(steekproef)
gemiddelde <-mean(steekproef)

afwijking
```

```
## [1] 15.98596
```

```
gemiddelde
```

```
## [1] 18.13753
```

B.

hoe groter de steekproef hoe dichter het gemiddelde zal liggen bij de 20. Hoeg groter de steekproef hoe duidelijker zal blijken dat deze collectie normaal verdeeld is

C.

```
zscore = (15.5-gemiddelde)/ afwijking
zscore
```

```
## [1] -0.1649905
```

```
zscore = (23-gemiddelde)/ afwijking
zscore
```

```
## [1] 0.3041711
```

D.

```
1 - pnorm(16, 20, 16)
```

```
## [1] 0.5987063
```

E.

```
pnorm(23, 20, 16)
```

```
## [1] 0.5743657
```

F.

```
ondergrens <- pnorm(16, 20, 16)
bovengrens <- pnorm(22, 20, 16)

bovengrens - ondergrens
```

```
## [1] 0.1484446
```

oefening 4.12.

examen vraag

opgave

Verkeersdrempels zijn bedoeld om de snelheid van automobilisten te beïnvloeden. Afhankelijk van de gewenste snelheid in een straat worden de drempels steiler of minder steil gemaakt. Drempel A is zo ontworpen dat 85 % van de automobilisten de drempel passeert met een snelheid van minder dan 50 km per uur. In de praktijk blijkt dat de passeersnelheid bij een drempel normaal verdeeld is. Bij drempel A werd een gemiddelde passeersnelheid van 43,1 km/h gevonden met standaardafwijking 6,6 km/h. 1. Toon aan dat 85% van de automobilisten niet harder dan 50 km/h rijdt. 2. Bij hoeveel van de 1200 metingen kan, op grond van eerdere ervaringen, een snelheid van meer dan 55 km/h worden verwacht?

oplossing

- cursusblad

a.

manier 1 –> qnorm: hoeveel procent zit er onder een bepaalde grens p (p is hier 0.85)

```
snelheid <- qnorm(mean = 43.1, sd = 6.6, p = 0.85)
snelheid
```

```
## [1] 49.94046
```

manier 2 –> pnorm: de cumulatieve kanfunctie voor een bepaalde waargenomen waarde x (x is hier dus 50)

```
snelheid <- pnorm(50, sd = 6.6, mean = 43.1)
snelheid
```

```
## [1] 0.8520935
```

manier 3 –> z-score: hoe ver ligt de geobserveerde waarde (50) van het gemiddelde

```
zscore <- (50 - 43.1) /6.6
zscore
```

```
## [1] 1.045455
```

als we gaan kijken in een z tabel vinden we dat de bijhorende kans 0.8531 is bij een alpha van 0.05

bron: http://www.z-table.com/

b.

```
n <- 1200
prob <- 1 - pnorm(mean = 43.1, sd = 6.6, q = 55)
prob
```

```
## [1] 0.03569173
```

```
aantal <- n * prob
aantal
```

```
## [1] 42.83007
```

oefening 4.13.

opgave

Gegeven 20 examenresultaten in Tabel 4.5. Uit resultaten van de laatste jaren blijkt dat $\sigma = 2.45$. 1. Wat is σ x , de standaardafwijking van x? 2. Geef het 92% betrouwbaarheidsinterval voor µ. 3. Kunnen we er zeker van zijn dat het gemiddeld resultaat minder dan 12.5 bedraagt?

oplossing

data

```
tabel <- c(11.5,16.5,11,17.3,10.8,5.6,13.1,11.5,14.2,12.9,8.7,9.2,15,14.4,10,10.3,18.3,12.9,14.2,8.7)
```

a.

gemiddelde en afwijking

```
mean <- mean(tabel)
mean
```

```
## [1] 12.305
```

```
deveation <- sd(tabel);
deveation
```

```
## [1] 3.194152
```

de standaardafwijking van de steekproef en de standaardafwijking van de populatie zijn niet het zelfde.

De centrale limietstelling zorgt er wel voor dat deze elkaar benaderen. Hoe groter n(steekproef) hoe dichter deze bij elkaar zullen liggen

b.

```
n <- length(tabel)

sd <- 2.45 / sqrt(20)

t <- qt(p = 0.04, df = n - 1)
# 0.04 * 2 = 0.08
# 1-0.08 = 0.92
# 0.92 = 92% betrouwbaarheidsinterval
t <- -t

links <- mean - t * sd
rechts <- mean + t * sd

links
```

```
## [1] 11.29176
```

```
rechts
```

```
## [1] 13.31824
```

alternatief: (scripts) in te geven data bij script:

```
data <- c(11.5, 16.5 ,11, 17.3, 10.8, 5.6, 13.1, 11.5, 14.2, 12.9, 8.7, 9.2, 15, 14.4, 10, 10.3, 18.3,

#data =  rnorm(100, 90, 60)

steekproefgemiddelde <- mean(data)
standaardAfwijkingPopulatie <- 2.45 # Hier NIET de standaardafwijking van de steekproef invullen!!!
n <- 20      # steekproefgrootte
betrouwbaarheidsinterval = 0.92 # Hier het betrouwbaarheidsinterval invullen! Niet alpha (wordt bereken
```

vervolg zie script “4. steekproefonderzoek/ betrouwbaarheidsinterval.R”

c.

gemiddelde van de steekproef is 12.305 linker grens is 11.29 rechter grens is 13.31

oefening 4.14.

(zelf gemaakt, niet de oplossing)

gegevens:

```
populatie <- 500
populatieGemiddelde <- 26
x <- 30
alpha <- 5
enkelzijdigeAlpha <- 2.5
```

oplossing:

??????????? moet standaardafwijking niet gegeven zijn?

oefening 4.15.

opgave

Een conservenfabrikant krijgt de laatste tijd klachten over de netto inhoud van zijn conserven met wortelen en erwtjes, die volgens de verpakking netto 1 liter zouden moeten bevatten. Daarom laat hij een steekproef nemen waarin de netto inhoud van 40 willekeurig gekozen blikjes wordt gecontroleerd. De resultaten worden samengevat in Tabel 4.6

vraag a: * Vul de tabel aan met de cumulatieve absolute frequentie * Vul de tabel aan met de relatieve frequentie * Vul de tabel aan met de cumulatieve relatieve frequentie.

vraag b:

- Bereken gemiddelde
- bereken de standaardafwijking
- hoeveel procent van de blikken bevatten te weinig wortelen en erwtjes
- teken een histogram van de absolute frequentie
- zijn de gegevens normaal verdeeld? hoe zie je dat?

oplossing

overnemen van de tabel

```
inhoud <- c('[970, 980[','[980, 990[','[990, 1000[','[1000, 1010[','[1010, 1020[','[1030, 1040[')

voorkomens <- c(975,975,975,985,985,985,985,985,995,995,995,995,995,995,995,995,995,995,995,995,995,1005
```

vraag a

hoe doe je dit met R????

vraag b

gemiddelde:

```
gem <- mean(voorkomens)
gem
```

```
## [1] 999.75
```

standaardafwijking:

```r
afwijking <- sd(voorkomens)
afwijking
```

```
## [1] 13.00641
```

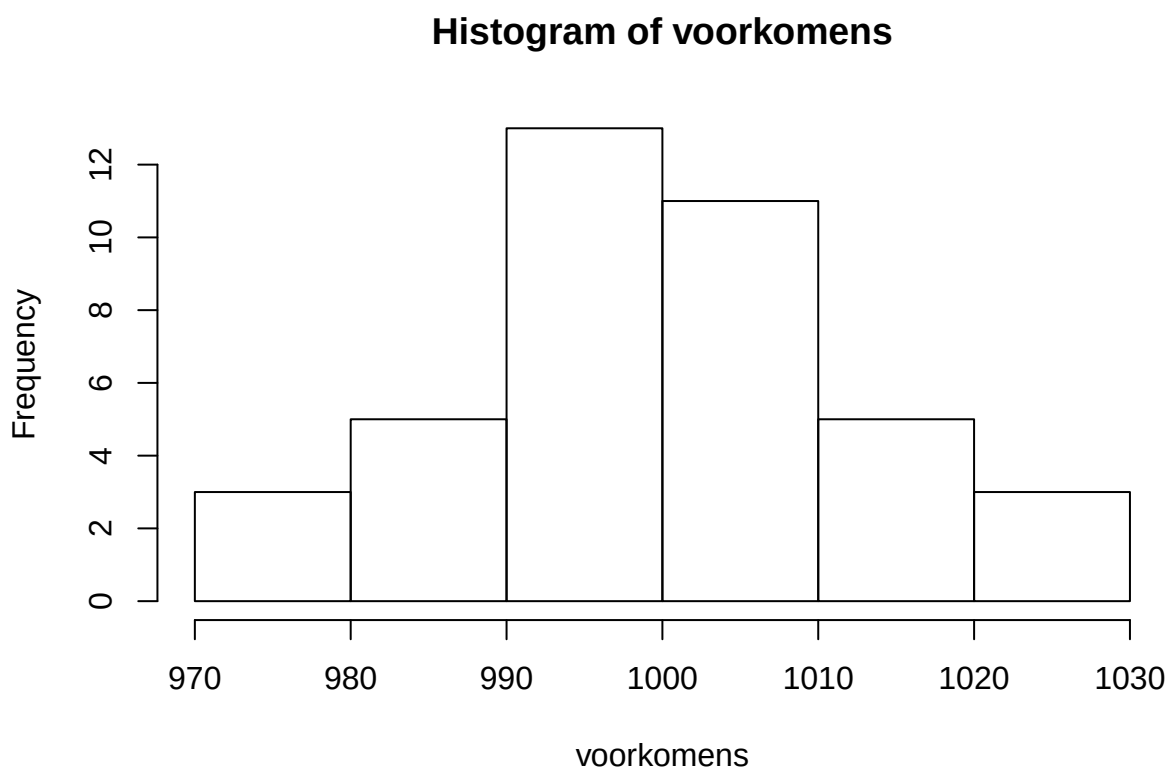
procent te kort:

```r
tekort <- pnorm(999, gem, afwijking)
tekort * 100
```

```
## [1] 47.70082
```

historgram:

```r
hist(voorkomens, breaks = 5)
```



normaal verdeeld?

antwoord

uit de histram kunnen we afleiden dat een normaal verdeling aanwezig is * klokvormig * symetrisch rond het gemiddelde

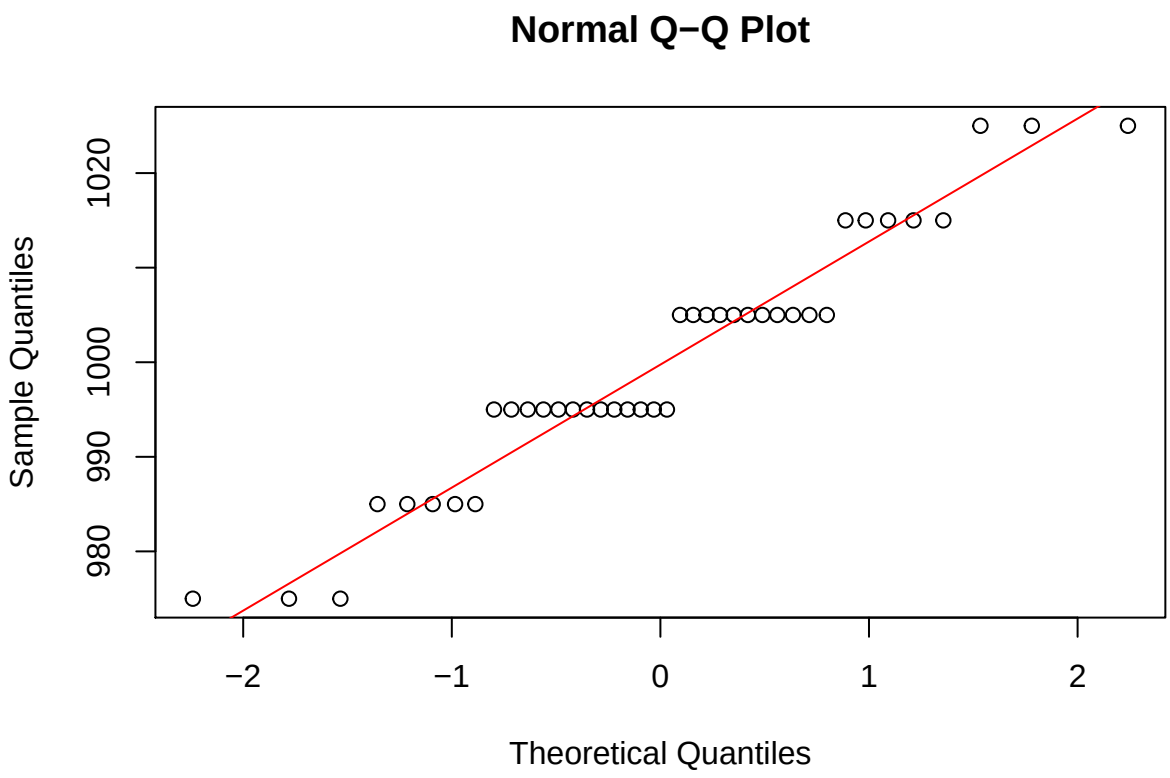
alternatieve manieren om dit te weten

1. berekenen van het aantal waarden binnen een afstand van 1 standaardafwijking

```r
# ?????
```

2. een qq plot maken

```
n <- length(voorkomens)
sig <- sd(voorkomens)
avg <- mean(voorkomens)

## bolletjes plaatsen
qqnorm(voorkomens)

## lijn maken
x <- seq(-3, 3, length = n)
lines(x ,gem + sig * x, col = "red")
```



de waarden liggen mooi verdeeld rond de rechte. dit duidt op een normale verdeling

3. bereken de kurtosis

```
#install.packages("e1071")

library(e1071)
```

```
## Warning: package 'e1071' was built under R version 3.5.3
```

```
kurtosis(voorkomens)
```

```
## [1] -0.4952053
```

een kortosis die niet ver van 0 afwijkt duidt op een normale verdeling. * 0 kurosis = normaal verdeeld * negatieve kurtosis = vlakke distributie * positieve kurtosis = piekvormige distributie

4. skewness berekenen

```r
skewness(voorkomens)
```

```
## [1] 0.0533888
```

een skewness die dicht bij 0 is duidt op een normale verdeling * 0 skewness = normaal verdeeld * negatieve skewness = lange linkse kant * positieve skewness = lange rechtse kant

oefening 4.16.

opgave

Een webhostingfirma heeft een Service Level Agreement met een klant voor een gegarandeerde uptime van "five nines" (99,999%). Die wordt aan het einde van elk jaar gecontroleerd en als de minimale uptime niet gehaald wordt, moet de hostingfirma een boete betalen.

Om de uptime te meten, voert een monitoringsysteem elke minuut een HTTP GET / uit en controleert het resultaat a.h.v. de HTTP return code. In de maand januari is er één enkele HTTP request onsuccesvol geweest.

- Als deze trend zich voortzet, wat is de kans dat de SLA niet gehaald wordt aan het einde van het jaar? Gebruik de formule voor de kansverdeling van een fractie.
- De gebruikte formule is eigenlijk niet geschikt in dit specifieke geval en geeft een vertekend beeld. Wat zou de reden kunnen zijn?

oplossing

```r
n <- 60*24*31
n
```

```
## [1] 44640
```

```r
# n = aantal minuten in de maand januari

aantalFalen <- 1
aantalSuccessen <- n - aantalFalen

succespercentage <- (aantalSuccessen/n) * 100
succespercentage
```

```
## [1] 99.99776
```

nee, als deze trend verder gezet wordt, zal de SLA niet gehaald worden.

Deze formule is niet geschikt voor dit soort vraagstukken omdat de fractie veel te klein is.