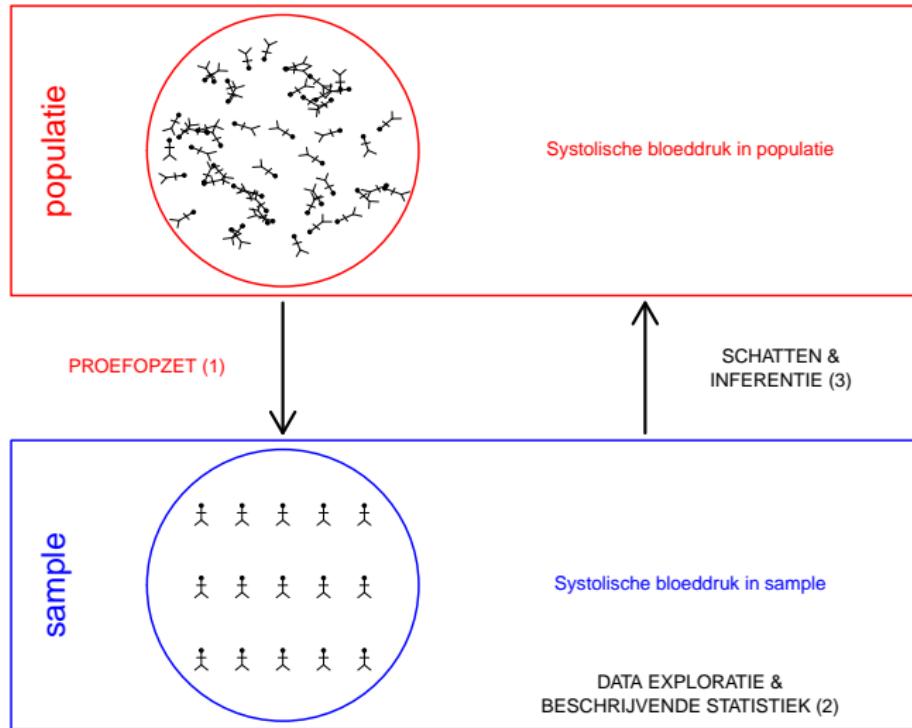


Hoofdstuk 4. Data exploratie

Lieven Clement

2^{de} bach. in de Biologie, Chemie, Biochemie en Biotechnologie en Biomedische Wetenschappen

4.1. Inleiding



- Raporteren van resultaten: niet zinvol om alle gegevens voor elk subject neer te schrijven
- Samenvatten en gericht visualiseren
- Inzicht in de data verwerven
- Fouten, anomalieën of zelfs fraude opsporen
- Veronderstellingen nagaan bv zijn de gegevens normaal verdeeld?

De eerste vraag die moet gesteld worden bij het benaderen van een echte data set is:

- ① Oorspronkelijke vraagstelling, waarom zijn deze gegevens verzameld?
- ② Hoe en onder welke omstandigheden zijn de subjecten gekozen en de variabelen gemeten? Design, aantal geplande subjecten, aantal geobserveerde objecten. Afhankelijkheid tussen subjecten?
- ③ Specifieke numerieke code voor ontbrekend gegeven of ander type uitzondering i.p.v. een echte meetwaarde?

Centrale dataset: de NHANES studie.

- Gegevens van 2009-2012 bij 10000 Amerikanen
- Ontbrekende waarnemingen code NA (Not Available / Missing Value)

| ID | Gender | Age | Race1 | Weight | Height | BMI | BPSysAve |
|-------|--------|-----|-------|--------|--------|-------|----------|
| 51624 | male | 34 | White | 87.4 | 164.7 | 32.22 | 113 |
| 51625 | male | 4 | Other | 17.0 | 105.4 | 15.30 | NA |
| 51630 | female | 49 | White | 86.7 | 168.4 | 30.57 | 112 |
| 51638 | male | 9 | White | 29.8 | 133.1 | 16.82 | 86 |
| 51646 | male | 8 | White | 35.2 | 130.6 | 20.64 | 107 |
| 51647 | female | 45 | White | 75.7 | 166.7 | 27.24 | 118 |

4.2. Univariate beschrijving van de variabelen

- Starten met *univariate* inspectie: elke variabele apart
- Via grafieken alvorens samenvattingssmaten (bv gemiddelde)
- Inzicht krijgen over verdeling van veranderlijke en of er eventuele *uitschieters* (d.i. extreme metingen of *outliers*) zijn.

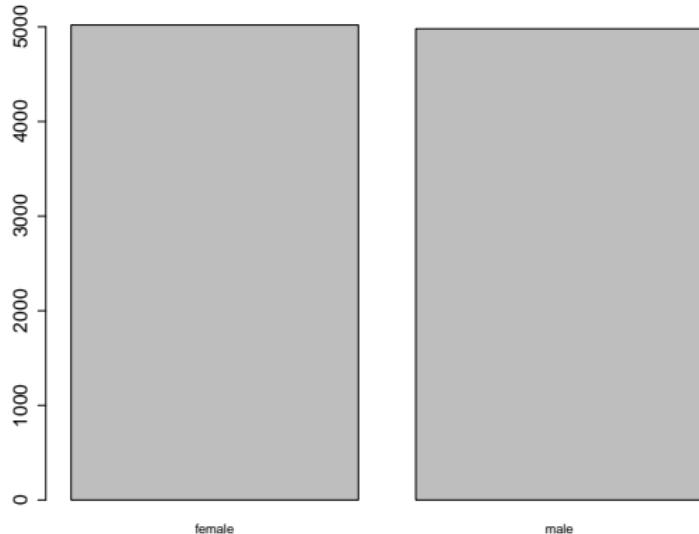
Nominale variabelen

- weinig methoden om te beschrijven
- Gender: kwalitatief nominaal

```
library(NHANES) #laad NHANES package
tab <- table(NHANES$Gender)
tab
```

```
##
## female    male
## 5020    4980
```

```
par(mar=c(5, 4, 4, 2) + 0.1,mai=c(1.02,0.82,0.82,0.42))  
barplot(tab,cex.lab=1.5,cex.axis=1.5,cex.main=1.5) #teken staaf
```



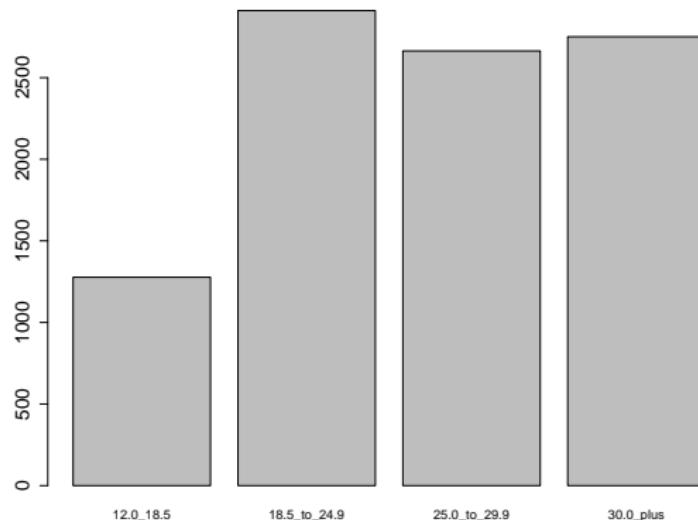
- Keuze tussen *absolute frequentie* (5020 voor het aantal vrouwen, 4980 voor het aantal mannen) of de *relatieve frequentie* (50.2% vrouwen, 49.8% mannen).

- *BMI_WHO* is kwalitatief ordinaal
- Ondergewicht, normaal gewicht, licht-overgewicht, obesitas
- Sorteren in volgorde
- Naast relatieve frequentie ook zinvol om weer te geven in *cumulatieve (relatieve) frequentie*: percentage van gegevens in gegeven klasse of lagere klasse

```
# sla freq. tabel op in object 'tabBmi'  
tabBmi <- table(NHANES$BMI_WHO)  
tabBmi  
  
##  
##      12.0_18.5 18.5_to_24.9 25.0_to_29.9      30.0_plus  
##          1277          2911          2664          2751
```

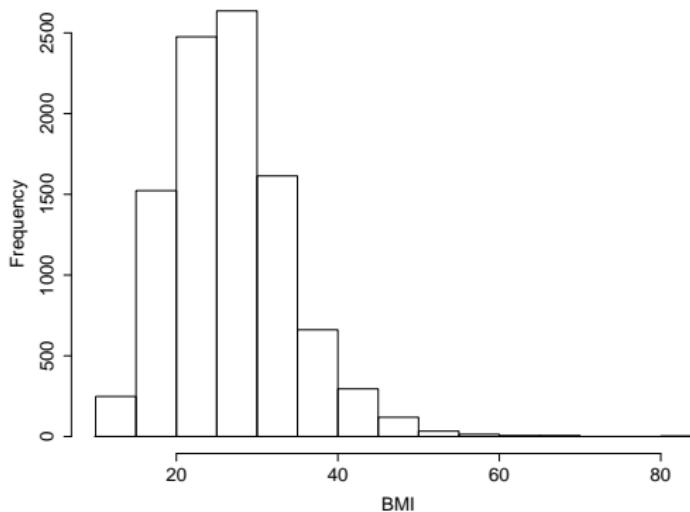
#teken staaf diagram

```
barplot(tabBmi,cex.lab=1.5,cex.axis=1.5,cex.main=1.5)
```



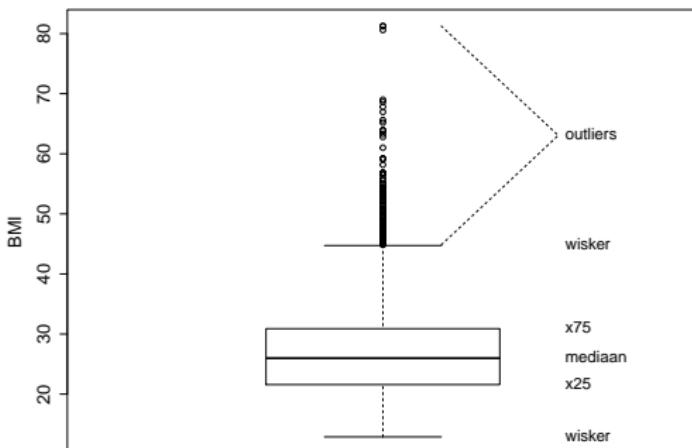
numerieke continue variabelen (histogram)

```
hist(NHANES$BMI,main="",xlab="BMI",cex.lab=1.5,cex.axis=1.5,cex.l  
ines(density(NHANES$BMI,na.rm=TRUE),main="",xlab="BMI")
```



Figuur 1: Histogram van het BMI in de NHANES studie.

Box en whisker plot



4.3. Samenvattingsmaten voor continue variabelen

4.3.1. Maten voor de centrale ligging

- Rekenkundig gemiddelde

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

```
mean(NHANES$BMI,na.rm=TRUE)
```

```
## [1] 26.66014
```

- Mediaan of 50% percentiel

```
median(NHANES$BMI,na.rm=TRUE)
```

```
## [1] 25.98
```



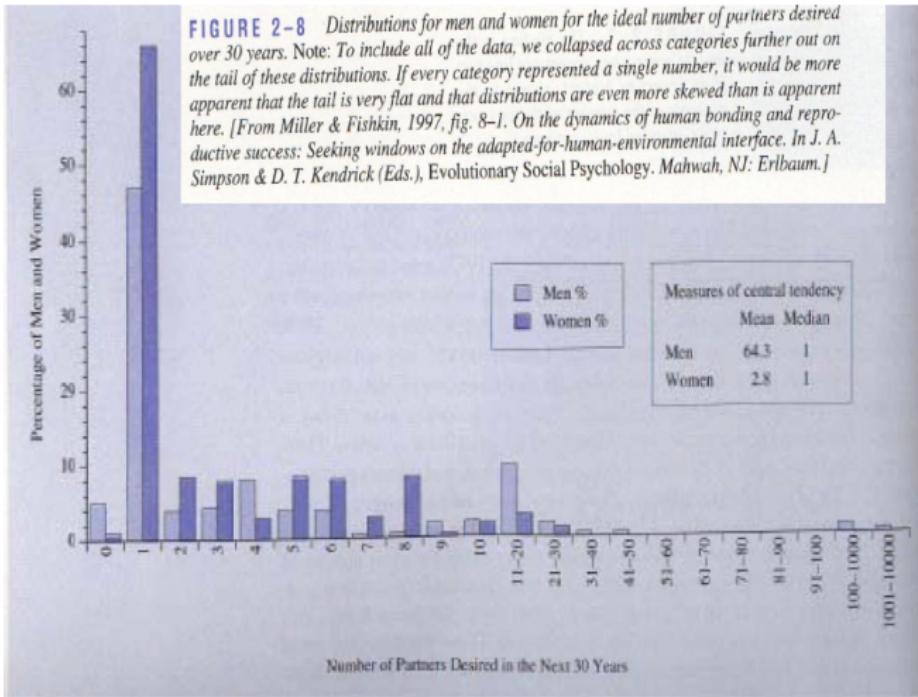
Mean of Mediaan?

- In een periode van 30 jaar, hopen mannen gemiddeld 64.3 partners te hebben, en vrouwen 2.8. (Miller and Fishkin, 1997)

Mean of Mediaan?

- In een periode van 30 jaar, hopen mannen gemiddeld 64.3 partners te hebben, en vrouwen 2.8. (Miller and Fishkin, 1997)
- In een periode van 30 jaar, is de mediaan van het aantal gewenste partners 1 bij zowel mannen als vrouwen. { (Miller and Fishkin, 1997)}
- Gemiddelde zeer gevoelig aan uitbijters!

FIGURE 2-8 Distributions for men and women for the ideal number of partners desired over 30 years. Note: To include all of the data, we collapsed across categories further out on the tail of these distributions. If every category represented a single number, it would be more apparent that the tail is very flat and that distributions are even more skewed than is apparent here. [From Miller & Fishkin, 1997, fig. 8-1. *On the dynamics of human bonding and reproductive success: Seeking windows on the adapted-for-human-environmental interface*. In J. A. Simpson & D. T. Kendrick (Eds.), *Evolutionary Social Psychology*. Mahwah, NJ: Erlbaum.]



Figuur 2: Partners

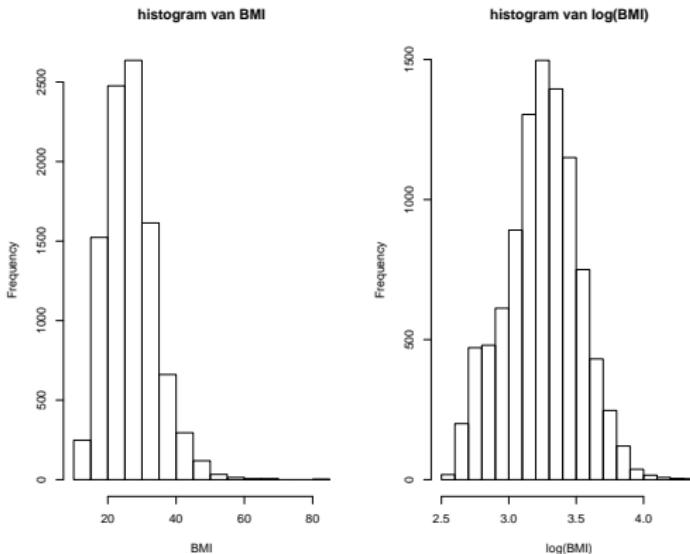
Geometrisch gemiddelde

Wortel aanpassen!!! -

$$\sqrt{\prod_{i=1}^n} = \exp \left\{ \frac{1}{n} \sum_{i=1}^n \log(x_i) \right\}$$

- Geometrisch gemiddelde dichter bij mediaan dan gemiddelde -
log-transformatie neemt scheefheid vaak weg
- Vaak een nuttigere maat voor centrale locatie dan mediaan:
 - 1 Gebruikt exacte waarden van alle observaties: meer precies algemene statistische technieken uit volgende hoofdstukken zijn direct toepasbaar voor een gemiddelde bv betrouwbaarheidsintervallen en hypothese toetsen
 - 2 Op een transformatie na is het een rekenkundig gemiddelde → algemene statistische technieken voor een gemiddelde zoals betrouwbaarheidsintervallen en hypothesesoetsen (zie volgende hoofdstukken) vrijwel rechtstreeks op geometrische gemiddelden.
 - 3 Vaak zinvol voor biologische gegevens zoals concentraties die niet

```
par(mfrow=c(1,2))
hist(NHANES$BMI, main="histogram van BMI", xlab="BMI")
hist(log(NHANES$BMI), main="histogram van log(BMI)", xlab="log(BM
```



Figuur 3: Boxplot van BMI en log(BMI) in de NHANES studie.

4.3.2. Spreidingsmaten

Spreiding van gegevens rond centrale waarde cruciaal:

- ➊ Om risico's te berekenen: centrale locatie + goede variëren gegevens
- ➋ Veldbiologen vaak geïnteresseerd in de mate waarin dieren of planten verspreid zijn in studiegebied.
- ➌ Vergelijking van uitkomsten: groepseffect duidelijker wanneer uitkomst weinig gespreid is dan wanneer gegevens meer chaotisch zijn. Spreiding is dus belangrijk om uit te maken of effecten toevallig zijn of systematisch.
- Uitkomsten variëren tussen individuen en binnen individuen: ligt aan basis van de statistische analyse
- Goed beschrijven van variatie naast centrale locaties is dus cruciaal.
- Welk deel van variatie kan men verklaren (door karakteristieken: behandeling, leeftijd, ...) en welk deel is onverklaard?

- Afwijkingen $x_i - \bar{x}$ zijn interessant om variatie in te schatten
- Gemiddelde afwijking is steeds 0
- Daarom gemiddelde kwadratische afwijkingen $(x_i - \bar{x})^2$ als maat voor variatie.
- Steekproef variantie:

$$S_x^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n - 1}$$

- Toevallig veranderlijke
- Als steekproefvariantie reeds geobserveerd wordt het als s^2 genoteerd

- Interpretatie vaak moeilijk: andere dimensie dan meetgegevens
- *Standaarddeviatie*:

$$s_x = \sqrt{s_x^2}$$

- Heel nuttig wanneer gegevens normaal verdeeld zijn:
- 68% van gegevens ligt tussen $\bar{x} - s_x$ en $\bar{x} + s_x$
- 95% van gegevens ligt tussen $\bar{x} - 2s_x$ en $\bar{x} + 2s_x$.
- Deze intervallen noemt men respectievelijk 68% en 95% *referentie-intervallen*.

```
sd(NHANES$BMI,na.rm=TRUE)
```

```
## [1] 7.376579
```

```
var(NHANES$BMI,na.rm=TRUE)
```

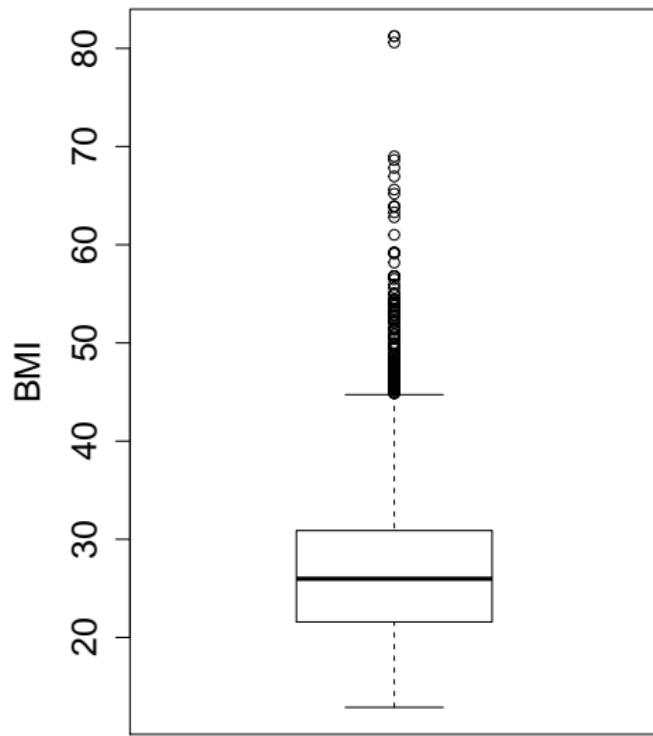
```
## [1] 54.41392
```

- Als gegevens niet normaal verdeeld zijn zijn de referentie-intervallen niet geldig
- Bij scheef verdeelde gegevens is standaarddeviatie niet langer interessant
- Bereik van gegevens gevoelig voor outliers.
- *Interkwartiel range*: afstand tussen derde kwartiel x_{75} en eerste quartiel x_{25} .
- Breedte boxplot!

```
IQR(NHANES$BMI,na.rm=TRUE)
```

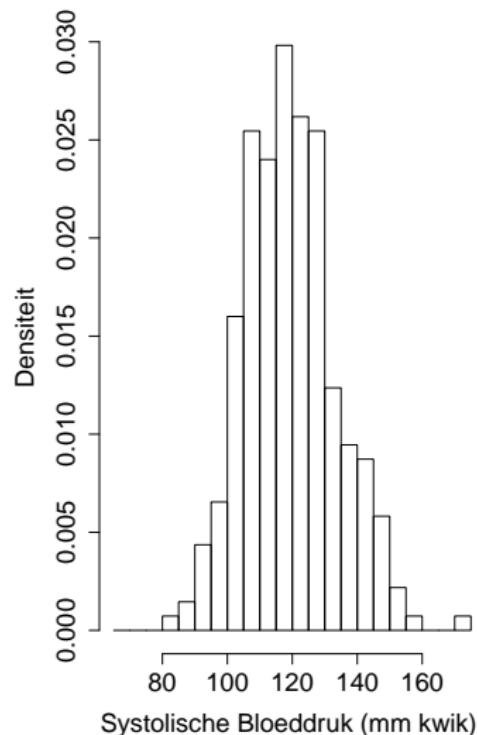
```
## [1] 9.31
```

```
boxplot(NHANES$BMI, ylab="BMI", cex.lab=1.5, cex.axis=1.5, cex.main=
```



4.4. De Normale benadering van gegevens

- Chapter 2 NHANES studie: Healthy subset.



- Biologische en chemische data vaak normaal verdeeld
- Dan kan je meer inzicht verwerven in gegevens a.d.h.v. minimaal aantal samenvattingsmaten: gemiddelde μ en standaard deviatie σ .
- *Normale curve of Normale dichtheidsfunctie:*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

- Geeft voor elke waarde x aan hoe frequent ze voorkomt.
- $\pi = 3.1459\dots$
- Symmetrisch rond gemiddelde
- Gemiddelde 0 en variantie 1: *standaardnormale curve of standaardnormale dichtheidsfunctie*.

- Oppervlakte onder normale curve: Kans of percentage.
- Veelal berekend via cumulatieve distributie $F(x)$:

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$$

- Normale dichtheidsfunctie zeer complex is, getal $F(x)$ niet expliciet uit te rekenen is
- Historisch standaard normale verdelingsfunctie getabuleerd.
- Standaard normale waarde z en $F(z)$ met $\Phi(z)$
- Symmetrie: $f(z) = f(-z)$ en $\Phi(-z) = 1 - \Phi(z)$
- $\Phi(-z)$: percentage dat kleiner is dan $-z$
- $1 - \Phi(z)$: het percentage dat groter is dan z .

- Vroeger ging men daarom eerst een normaal verdeelde variabele standardiseren om ze om te zetten naar standaard normaal verdeling:

$$Z = \frac{X - \mu}{\sigma}$$

- Aangezien voor een willekeurig getal x

$$X \leq x \Leftrightarrow \frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}$$

vinden we nu dat

$$\begin{aligned} P(a \leq X \leq b) &= P\left(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}\right) \\ &= \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right) \end{aligned}$$

Conventie

- z_α waarde aan waar $\alpha 100\%$ van de oppervlakte onder de standaardnormale curve rechts zit
- $P(Z \geq z_\alpha) = \alpha$.
- Voor $z_{\alpha/2}$ geldt dat $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$.
- Bijvoorbeeld, $P(-z_{0.025} \leq Z \leq z_{0.025}) = 95\%$. ✓
- Of $[-z_{\alpha/2}, z_{\alpha/2}]$ dus $(1 - \alpha)100\%$ van de observaties.

- Stel dat X een Normaal verdeelde meting is met gemiddelde μ en standaarddeviatie σ . Dan geldt dat

$$P\left(-z_{\alpha/2} \leq \frac{X - \mu}{\sigma} \leq z_{\alpha/2}\right) = 1 - \alpha$$

Hieruit volgt dat

$$P(\mu - z_{\alpha/2}\sigma \leq X \leq \mu + z_{\alpha/2}\sigma) = 1 - \alpha$$

- Voor Normaal verdeelde metingen met gemiddelde μ en standaarddeviatie σ bevat het interval $[\mu - z_{\alpha/2}\sigma, \mu + z_{\alpha/2}\sigma]$ dus $(1 - \alpha)100\%$ van de observaties.
- In de praktijk worden de parameters μ en σ hierbij vervangen door \bar{x} en s_x .
- **referentie-interval**
- Is bijvoorbeeld belangrijk voor opsporen van pathologie: e.g. hypertensie.

- De systolische bloeddruk voor “gezonde” personen is symmetrisch
- we zullen later aantonen dat deze approximatif normaal verdeeld zijn.
- In de gezonde subset is steekproefgemiddelde 119.5mmHg en standaarddeviatie 14.1mmHg.
- Als we het populatiegemiddelde en het populatiestandaarddeviatie door deze schattingen vervangen dan bekomen we volgend referentie interval: [91.9, 147]mmHg.

```
mean(nhanesSubHealthy$bpSys) + qnorm(c(0.025, 0.975)) * sd(nhanesSubH
```

```
## [1] 91.88971 147.03393
```

- Merk op dat we in hoofdstuk 2 hadden we gebruik gemaakt van een eenzijdig referentie-interval

4.4.2 QQ-plots

- Als men steunt op de normale verdeling is het belangrijk om na te gaan of aan deze aannames is voldaan.
- In deze cursus doen we dat a.d.h.v. *QQ-plots* of *kwantielgrafieken* (in het Engels: *quantile-quantile plots*) erdeelde gegevens.
- Percentielen van observaties in steekproef uitgezet t.o.v. overeenkomstige percentielen van Normale curve.
- Als gegevens Normaal verdeeld zijn, komen beide percentielen vrij goed overeen
- Punten van grafiek worden min of meer op rechte verwacht
- Systematische afwijkingen van een rechte wijzen op systematische afwijkingen van Normaliteit.
- Lukrake afwijkingen van een rechte kunnen gevolg zijn van toevallige biologische variatie en zijn daarom niet indicatief voor afwijkingen van Normaliteit.

- we doen dit eerst op basis van gesimuleerde data waarvoor we de verdeling kennen.

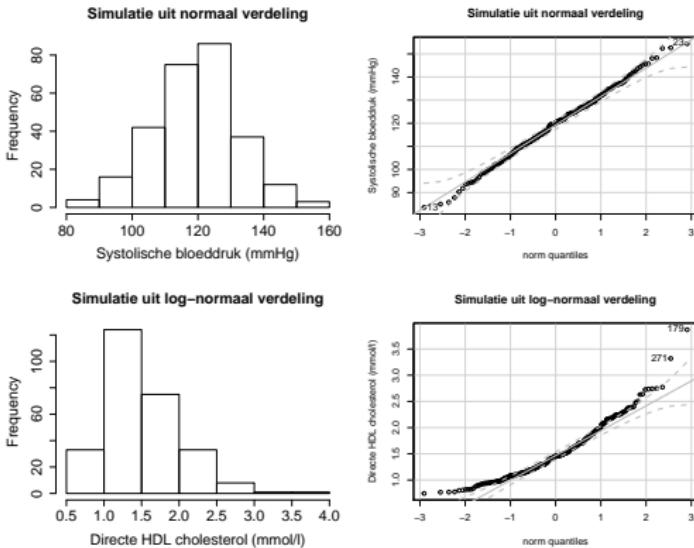
```
library("car")
simNorm=rnorm(nrow(nhanesSubHealthy),mean(nhanesSubHealthy$bpSys
hist(simNorm,xlab="Systolische bloeddruk (mmHg)",main="Simulatie
qqPlot(simNorm,ylab="Systolische bloeddruk (mmHg)",main="Simulat
```

```
## [1] 60 235
```

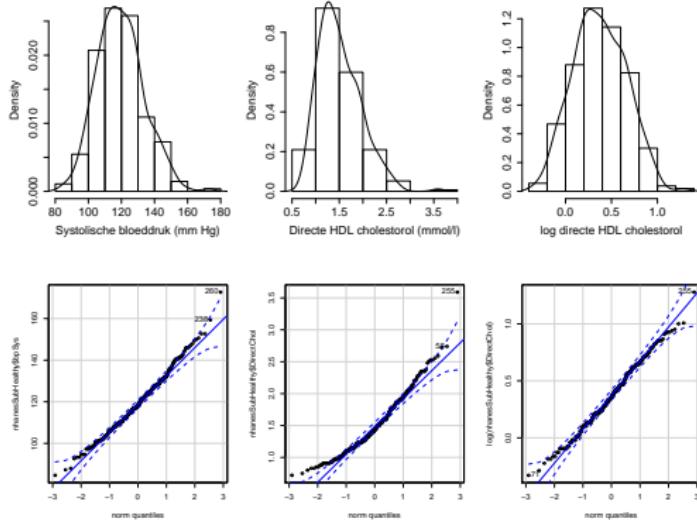
```
simLogNorm=exp(rnorm(nrow(nhanesSubHealthy),mean(log(nhanesSubHe
hist(simLogNorm,xlab="Directe HDL cholesterol (mmol/l)",main="Si
qqPlot(simLogNorm,ylab="Directe HDL cholesterol (mmol/l)",main="
```

```
## [1] 150 181
```





Echte data



4.5 Samenvattingsmaten voor categorische variabelen

We zullen deze sectie behandelen in hoofstuk 9 wanneer we dieper in gaan op de analyse van categorische data