UNIVERSITEIT GENT

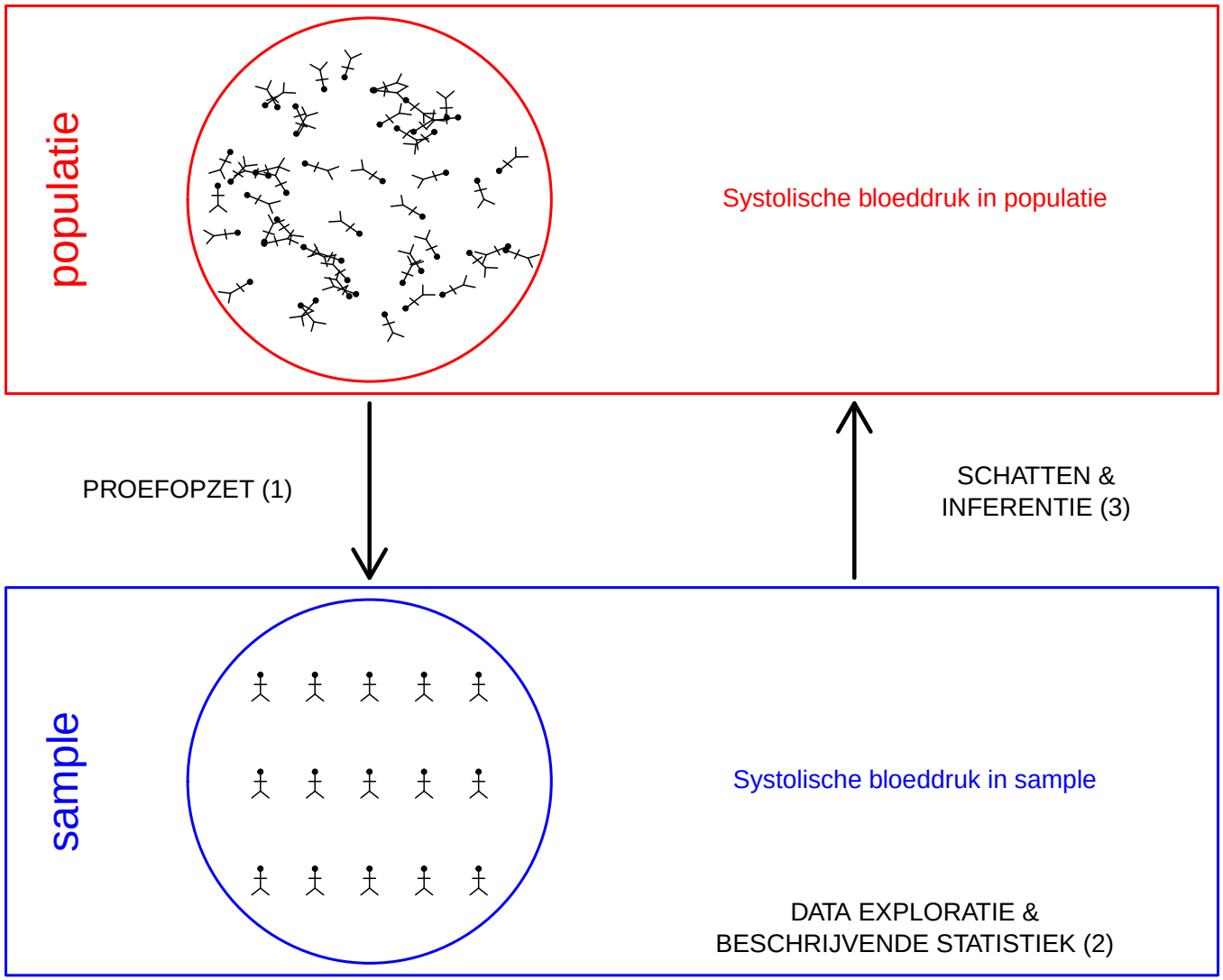
BIG N2N
Bioinformatics Institute Ghent
from nucleotides to networks

Belangrijke concepten & conventies

Lieven Clement

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

Inleiding


populatie
Systolische bloeddruk in populatie
PROEFOPZET (1)
SCHATTEN &
INFERENTIE (3)
sample
Systolische bloeddruk in sample
DATA EXPLORATIE &
BESCHRIJVENDE STATISTIEK (2)

1. **Proefopzet (1)**

   - Eerst bepaalt onderzoeker **populatie** van interesse.
   - Financiële en logistieke beperkingen $\rightarrow$ **representatieve steekproef** uit de populatie

2. Data-analyse

   - **Data-exploratie en beschrijvende statistiek (2)**: verkennen, visualiseren, samenvatten, inzicht, assumpties
   - **Statistische Besluitvorming of Inferentie (3)**: Wat we observeren in de steekproef trachten te veralgemenen naar de algemene populatie toe, zodat we algemene conclusies kunnen trekken op populatie-niveau op basis van de steekproef van de studie.

Voorbeeld

- National Health and Nutrition Examination Survey (NHANES)
- Amerikaanse demografische studie sinds 1960 op regelmatige basis afgenomen
- Groot aantal fysieke, demografische, nutritionele, levelsstijl en gezondheidskarakteristieken gecollecteerd in deze studie

	ID	Gender	Age	Race1	Weight	Height	BMI	BPSysAve
1	51624	male	34	White	87.4	164.7	32.22	113
4	51625	male	4	Other	17.0	105.4	15.30	NA
5	51630	female	49	White	86.7	168.4	30.57	112
6	51638	male	9	White	29.8	133.1	16.82	86
7	51646	male	8	White	35.2	130.6	20.64	107
8	51647	female	45	White	75.7	166.7	27.24	118

Variabelen

- We meten *variabelen* op de subjecten in de steekproef
- Variabele is een karakteristiek bvb. Systolische bloeddruk, leeftijd, geslacht, ...
- Varieert van subject tot subject in de studie.

Types variabelen

1. *Kwalitatieve variabelen* beperkt aantal uitkomstcategoriën, niet numeriek
   - *nominale variabelen*: geen natuurlijke ordening, vb geslacht, ras, bloedgroep, kleur van ogen...
   - *ordinale variabelen*: wel een ordening, vb BMI klasse volgens WHO, rokersstatus (1: nooit gerookt, 2: ooit gerookt maar gestopt, 3: actueel roker)

   Vaak numerieke rangen maar meestal hebben deze geen betekenis!
2. *Numerieke variabelen*:
   - *discrete variabelen*: tellingen vb aantal partners die men had gedurende het leven, ...
   - *continue variabelen*: kunnen (tenminste in theorie) tussen bepaalde grenzen elke mogelijke waarde aannemen vb leeftijd, gewicht, BMI, fluorescentie-meting in ELISA
   - Dichotomiseren om nominaal te maken $\rightarrow$ informatie verlies

Populatie

- Doel van wetenschappelijke studie: uitspraken doen over de algemene populatie.
- Vb grenswaarde afleiden om patiënten met hypertensie op te sporen

$\rightarrow$ systolische bloeddruk bestuderen bij een populatie van gezonde personen

- Populatie is theoretisch concept
  - Is meestal continu in verandering
  - Vaak ook interesse in toekomstige subjecten *rightarrow* dus op bepaald ogenblik niet volledig observeerbaar
  - kan als oneindig groot worden beschouwd
- Populatie duidelijk omschrijven!

Populatie duidelijk omschrijven

Inclusiecriteria zijn karakteristieken die een subject/experimentele eenheid moet hebben om tot de populatie te behoren, b.v.

- leeftijdscategorie 45-65
- normaal BMI
- ...

Exclusiecriteria zijn karakteristieken die een subject/experimentele eenheid niet mag hebben om tot de populatie te behoren, b.v.

- diabetes
- historiek van hard drugs
- lage gezondheidsstatus
- slaapproblemen
- ...

Toevalsveranderlijken (toevallige veranderlijken)

- Variabelen (vb Systolische bloeddruk) variëren in de populatie van subject tot subject!
- Variabelen zijn dus *random of veranderlijk* aangezien hun waarde veranderlijk is in de populatie
- **Cruciale vraag**: Hoe nauwkeurig zijn uitspraken over de populatie o.b.v. een groep gemeten subjecten in een steekproef!
- We zullen dus steeds verschillen zien van steekproef tot steekproef
- Spreiding op gegevens speelt cruciale rol

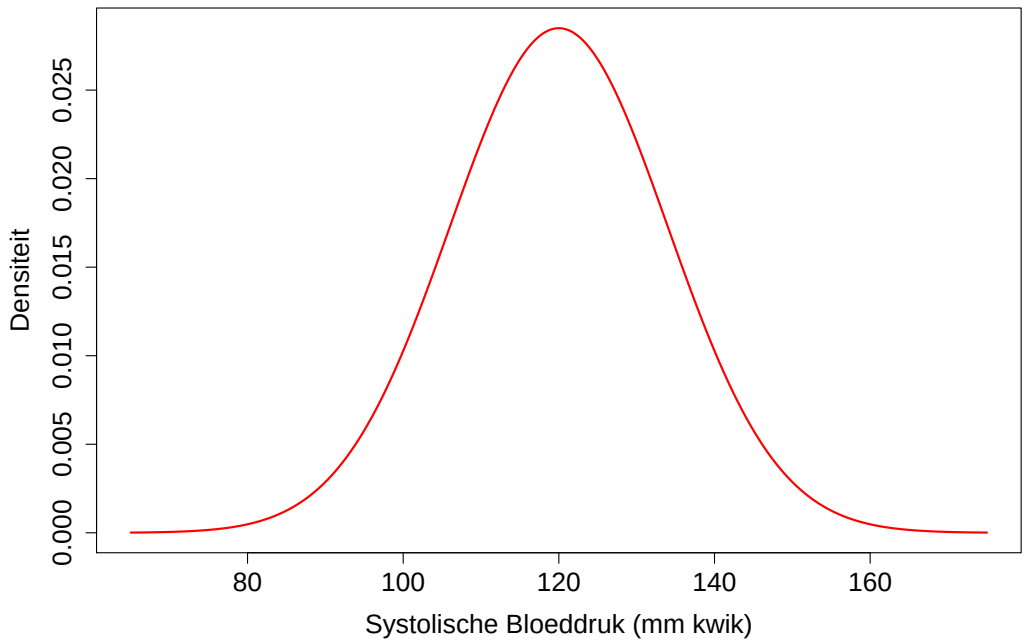
Conventie

- Gebruik hoofdletters om aan te geven dat bestudeerde karakteristiek (vb. systolische bloeddruk) variabel is in de populatie zonder daarbij concreet over de gerealiseerde waarde voor een bepaald subject na te denken.
- Variabele X wordt algemeen een *toevalsveranderlijke* genoemd: is formeel resultaat van een *toevallige trekking* van een bepaalde karakteristiek uit de studiepopulatie.
- X is dus label van bepaalde populatiekarakteristiek voor een lukraak individu uit de bestudeerde populatie, vooraleer haar concrete waarde gemeten werd.
- Een toevalsveranderlijke X kan men dus opvatten als onbekende veranderlijke die een meting voorstelt die we plannen te verzamelen, maar nog niet hebben verzameld.
- Is noodzakelijk om hierover na te denken zodat we kunnen redeneren hoe resultaten van steekproef tot steekproef kunnen wijzigen
- Net zoals observaties kunnen we toevallig veranderlijken klasseren als kwalitatief, kwantitatief, discreet, continu, ...

Beschrijven van de populatie

- Voor we een random variabele meten is het onmogelijk te zeggen hoe hoog de meting precies zal zijn.
- Gerealiseerde waarde van X zijn dus onderhevig aan random variabiliteit
- Een toevalsveranderlijke X wordt beschreven door gebruik te maken van een *verdeling*.
- De verdeling beschrijft waarschijnlijkheid om bepaalde waarde te observeren voor toevallig veranderlijke X wanneer men lukraak een proefpersoon kiest uit de populatie.
- De densiteitsfunctie van de verdeling wordt vaak genoteerd als f(x).
- Veel biologische karakteristieken volgen Normale verdeling: $f(x) = N(\mu, \sigma^2)$.
- Stel dat de gemiddelde bloeddruk $\mu = 120$ mmHg en variantie $\sigma^2 = 196$ in populatie.

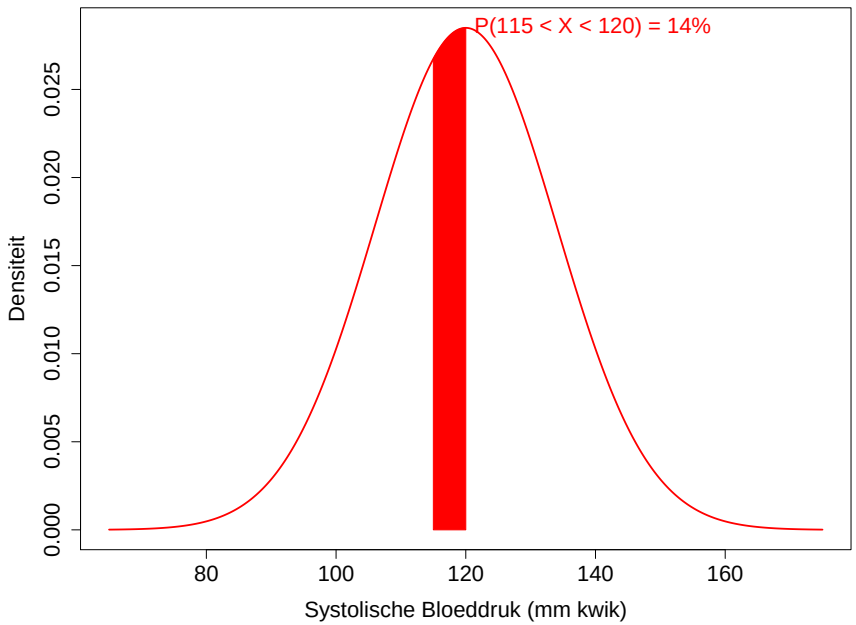
```r
par(mar=c(5, 4, 4, 2) + 0.1,mai=c(1.02,0.82,0.82,0.42))
grid <- seq(65,175,.1)
plot(grid,dnorm(grid,mean=120,sd=196^.5),
     xlab="Systolische Bloeddruk (mm kwik)",
     col=2,ylab="Densiteit",type="l",lwd=2,cex.lab=1.5,cex.axis=
```



Kans om dat lukraak individu uit de populatie te trekken met bloeddruk tussen 115 en 120 mmHg

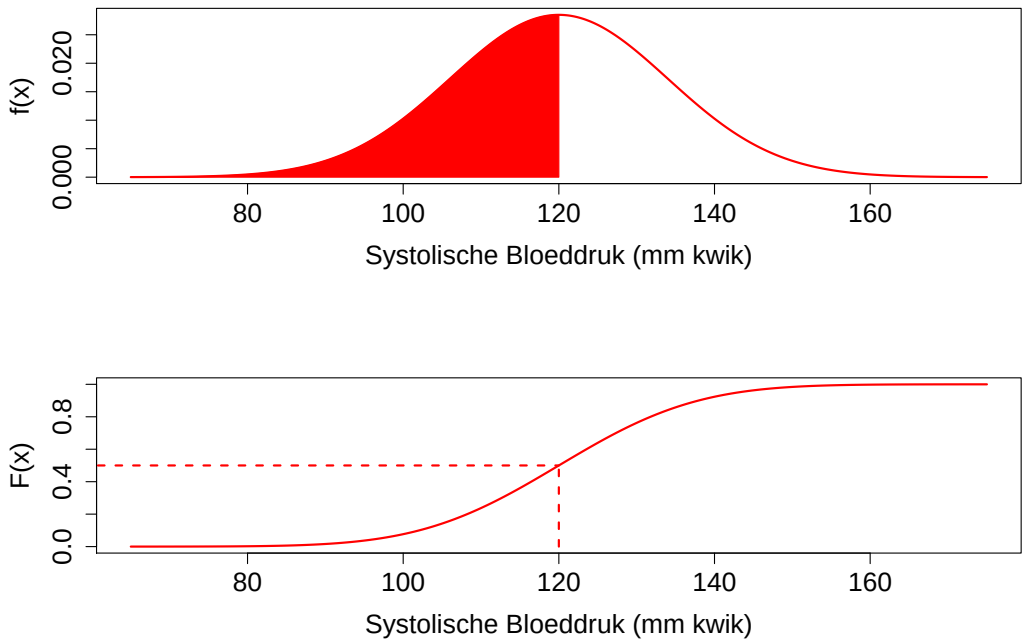
$$P[115 \leq X \leq 120] = \int_{x=115}^{120} f(x)dx$$

Oppervlakte onder de densiteitsfunctie is 1!



Kansen worden meestal berekend door gebruik te maken van de *cumulatieve distributie functie* van de verdeling:

$$F(x) = \int_{-\infty}^{x} f(x)dx$$



- In R a.d.h.v. de functie `pnorm()`.
- Op basis van de Normaal verdeling voor de systolische bloeddruk in de populatie $N(\mu = 120, \sigma^2 = 196)$ bekomen we

$$P(115 \leq X \leq 120) = F(120) - F(115)$$

- Merk op dat de normale distributie in R geparameteriseerd wordt a.d.h.v. het gemiddelde en de standaard afwijking

```
pnorm(120,mean=120,sd=196^.5)-pnorm(115,mean=120,sd=196^.5)
```

```
## [1] 0.1395076
```

Steekproef

- In de praktijk kennen we de werkelijke verdeling in de populatie niet
- Om financiële en logistieke redenen bijna nooit mogelijk om volledige populatie te bestuderen.
- Populatieparameters (vb gemiddelde bloeddruk, variantie van bloeddruk) kunnen daarom meestal niet exact bepaald worden.
- Enkel een deel van de populatie kan onderzocht worden, hetgeen men de *steekproef* noemt.
- Trekken volgens gestructureerd design worden daartoe **lukraak subjecten** uit de doelpopulatie getrokken en geobserveerd: **Representativiteit**.
- De onbekende populatieparameters worden vervolgens geschat o.b.v. die steekproef en noemt met schattingen.
- De steekproef $x_1, x_2, ..., x_n$ kan als n realisaties worden beschouwd van dezelfde toevallige veranderlijke X, voor subject i, met $i = 1, 2, ..., n$.

NHANES voorbeeld

- O.b.v. een studie hypertensie definiëren.
- Subjecten met "normale" bloeddrukwaarden selecteren :inclusie-exclusie criteria
- Selectie van n gezonde personen tussen 40 en 65 jaar en interesse in systolische bloeddruk.
- Eens lukraak individu getrokken wordt uit populatie heeft men realisaties van de toevalsveranderlijke X kunnen observeren.
- *Conventie*: Geobserveerde waarde duiden we aan met een kleine letter x.
- x is een welbepaald getal en niet langer een onbekende veranderlijke
- Samengevat: De nog onbekende waarden voor de bestudeerde populatiekarakteristiek bij subjecten 1 tot n in de steekproef zijn toevalsveranderlijken: $X_1, ..., X_n$.
- Na het trekken van de steekproef ziet men de gerealiseerde uitkomsten $x_1, x_2, \ldots, x_n$, bijvoorbeeld hun gemeten systolische bloeddruk.

Populatie van gezonde personen tussen de 40-65 jaar: inclusie & exclusie-criteria

- $n = 275$ gezonde individuen uit de Amerikaanse populatie weerhouden (uit NHANES studie).

```
library(NHANES)
NHANES2=subset(NHANES,!is.na(Race1)&!is.na(Smoke100n)&!is.na(BMI
NHANES2$bpSys=rowMeans(NHANES2[,c(27,29,31)])
nhanesSub=subset(NHANES2, Age<=65&Age>=40 &!duplicated(ID) )
nhanesSubHealthy=subset(nhanesSub,Smoke100n=="Non-Smoker"&Diabet
head(nhanesSubHealthy$bpSys)
```

```
## [1] 114.00000 140.66667  94.66667 152.66667 128.00000 124.000
```

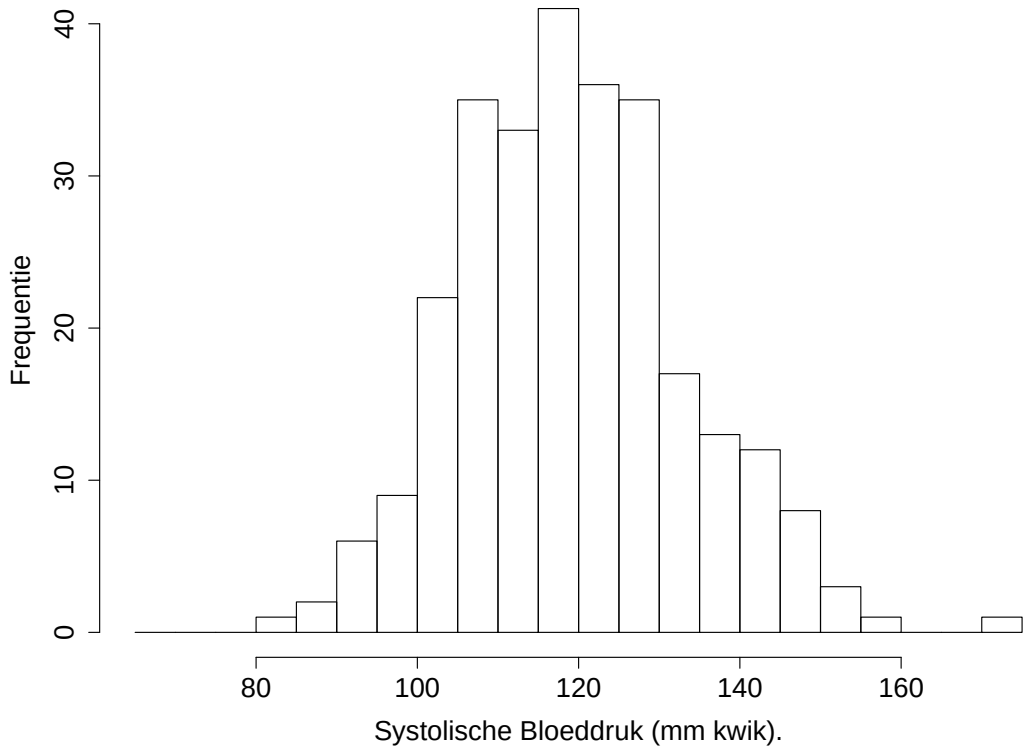
```
dim(nhanesSubHealthy)
```

```
## [1] 275  77
```

Schatten van de verdeling in de populatie

- Verdeling o.b.v. steekproef schatten a.d.h.v. een histogram

```
histAbs<-hist(nhanesSubHealthy$bpSys,
                xlab="Systolische Bloeddruk (mm kwik).",
          breaks=seq(65,175,5),ylab="Frequentie",cex.main=1.5,cex.
```



- Kans schatten dat random persoon in populatie bloeddruk tussen 115 - 120 mm Hg heeft.

```
tab<-cbind(histAbs$mids,histAbs$counts)
head(tab)
```

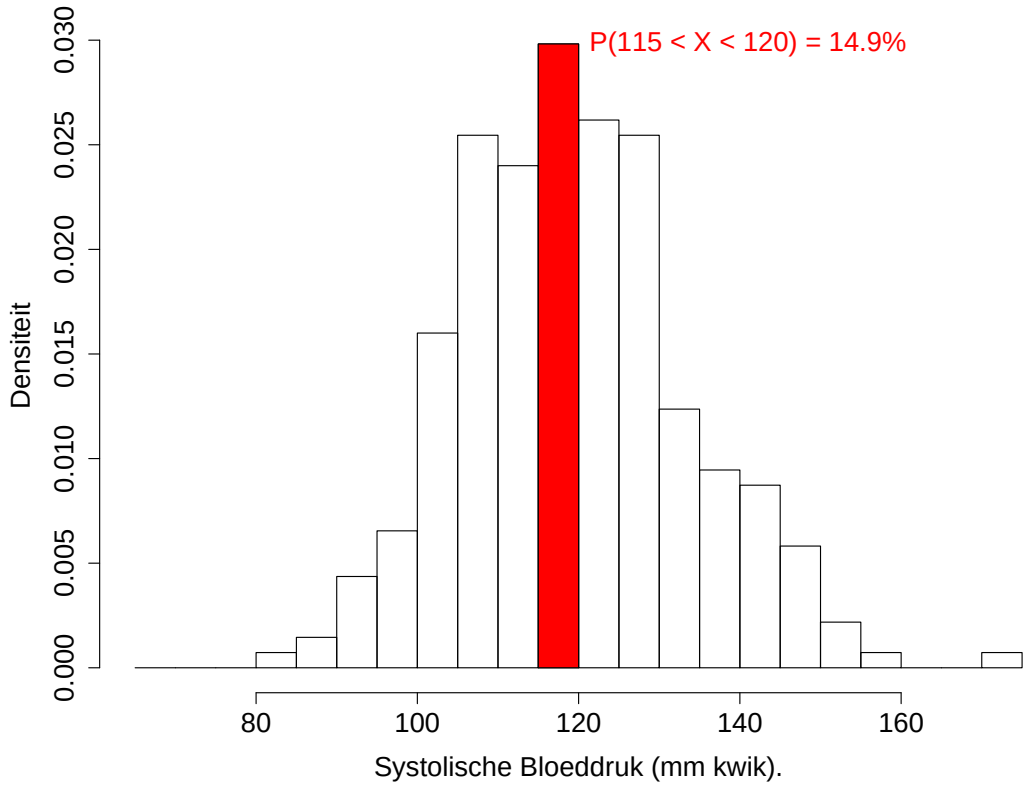
```
##       [,1] [,2]
## [1,] 67.5    0
## [2,] 72.5    0
## [3,] 77.5    0
## [4,] 82.5    1
## [5,] 87.5    2
## [6,] 92.5    6
```

```
tab[tab[,1]==117.5,2]/sum(tab[,2])
```

```
## [1] 0.1490909
```

Een histogram met relatieve frequenties/densiteiten.

```
hist(nhanesSubHealthy$bpSys,xlab="Systolische Bloeddruk (mm kwik
rect(115,0,120, tab2[tab2[,1]==117.5,2],col=2)
text(120,tab2[tab2[,1]==117.5,2],paste0("P(115 < X < 120) = ",ro
```



- Oppervlakte in elke balk komt overeen met kans

- Balkbreedte hier 5 mm Hg gekozen.

```
tab2<-cbind(histAbs$mids,histAbs$density)
tab2[tab2[,1]==117.5,2]
```

```
## [1] 0.02981818
```

```
tab2[tab2[,1]==117.5,2] * 5
```

```
## [1] 0.1490909
```

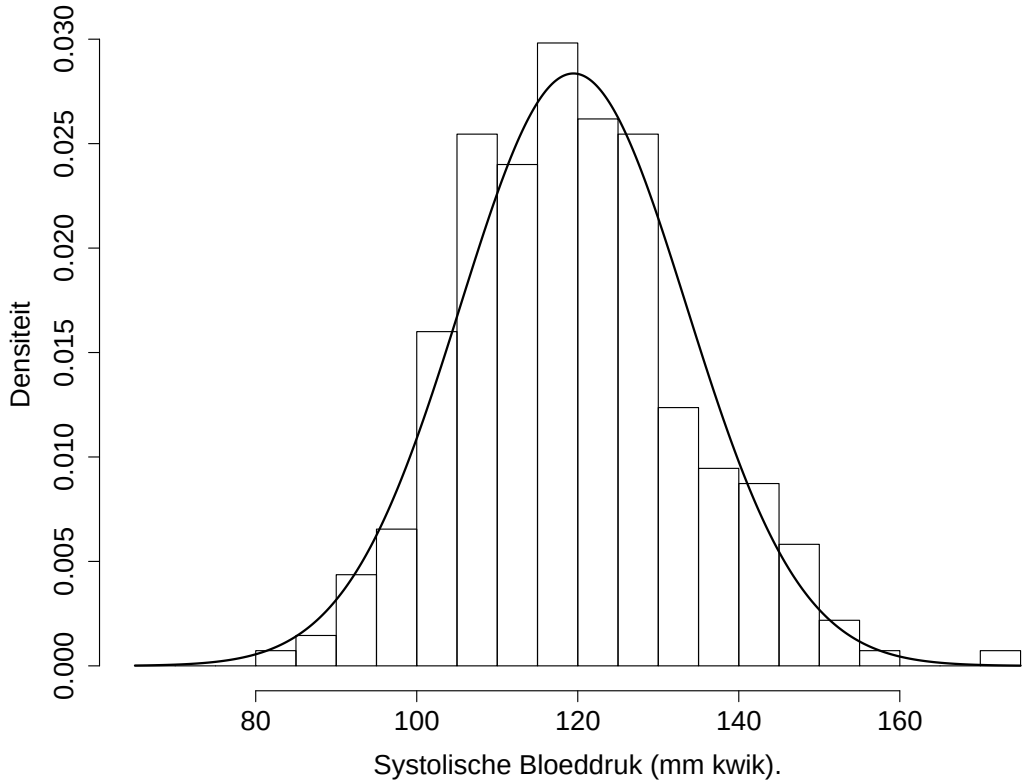
- Som van alle oppervlaktes = 1! Kans om willekeurig object te zien in de steekproef met bloeddruk tussen minimale en maximale bloeddruk die werd geobserveerd in de steekproef.

```
sum(tab2[,2]*5)
```

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

Aanname van normaliteit

- Ipv histogram kunnen we de verdeling ook schatten na aanname van normaliteit!
- Gebruik Normale verdeling met als gemiddelde en variantie de schattingen uit de steekproef



We kunnen ook opnieuw kansen berekenen o.b.v. geschatte normale verdeling.

```
xBar <- mean(nhanesSubHealthy$bpSys)
sBar <- sd(nhanesSubHealthy$bpSys)
pnorm(120,mean=xBar,sd=sBar)-pnorm(115,mean=xBar,sd=sBar)
```

```
## [1] 0.1397006
```

- Schatting ligt veel dichter bij werkelijke kans in populatie!
- Nauwkeuriger dan histogram: alle waarden gebruikt om gemiddelde en standaard afwijking te schatten
- Histogram: enkel waarden tussen 115 en 120!
- Wel aannames nagaan: is één van de doelstellingen van data exploratie.

Drempelwaarde voor hypertensie?

- O.b.v. steekproef en aannames van normaliteit kunnen grenswaarde afleiden die extreem is voor "gezonde" personen in de populatie.
- V.b. Bloeddruk te bepalen die maar met een kans van 5% wordt overschreden

$$P(X > t_{\text{drempel}}) = 5\%$$

of

$$P(X \leq t_{\text{drempel}}) = 95\%$$

- Dat kan met de functie `qnorm()` in R.

```
qnorm(0.95,mean=xBar,sd=sBar)
```

```
## [1] 142.6011
```

- Deze waarde ligt dicht bij 140 mmg Hg, een grenswaarde voor hypertensie die vaak in de literatuur wordt gebruikt.

Statistieken

- Formules die gebruikt worden om parameters van de verdeling in de populatie te schatten o.b.v. steekproef
- alsook numerieke resultaat dat men bekomt door deze formules te evalueren, worden *statistieken* genoemd.
- Bv. rekenkundig gemiddelde van alle systolische bloeddrukwaarden in de steekproef
- Onderzoekers willen ongekende parameters van populatie weten en schatten die met statistieken geobserveerd of berekend o.b.v. steekproef.
- Omdat statistieken berekend worden op basis van de gegevens uit de steekproef, zullen ze variëren van steekproef tot steekproef.
- Statistieken worden dus genoteerd met een hoofdletter (bvb. $\bar{X}$ voor het steekproefgemiddelde)
- Tenzij we verwijzen naar de numerieke waarde die gerealiseerd wordt in een bepaalde steekproef: dan een kleine letter bv. $\bar{x}$ voor het steekproefgemiddelde.

Belangrijke Conventie

- **Populatieparameters** die een vaste waarden aannemen maar die meestal ongekend zijn $\rightarrow$ **Griekse symbolen**.
- **Statistieken** waarmee we deze ongekende parameters schatten o.b.v. een steekproef $\rightarrow$ **letters**.
- Vb Normale verdeling

Populatie	Steekproef
μ	$\bar{X}$
σ^2	S^2