

Les 1: de normale distributie

Koen Van den Berge
Statistiek

2^e Bachelor in de biologie
2^e Bachelor in de chemie

16 oktober 2018

Indeling lessen

Elke bullet point is een week.

- ▶ R en normale distributie
- ▶ one-sample t-test
- ▶ two-sample t-test
- ▶ Regressie
- ▶ ANOVA
- ▶ niet-parametrische testen
- ▶ categorische data analyse

Elk practicum bestaat uit drie blokken: (i) volledig uitgewerkte analyse die overlopen wordt; (ii) samen tweede analyse op een andere of uitgebreide dataset; (iii) case study.

Indeling lessen: Case studies

- ▶ In totaal zullen jullie zes maal een aantal multiple choice vragen dienen op te lossen via Minerva.
- ▶ Uitgezonderd van de eerste week (vandaag), zullen deze elk meetellen voor één punt op het examen (→ 5/20 punten op case studies).
- ▶ Elke case study zal gebaseerd zijn op een werkelijk probleem, en de respectievelijke onderzoek(st)er zal haar/zijn probleem eerst aan jullie voorstellen adhv een presentatie.
- ▶ Jullie dienen de statistische analyse uit te voeren in een RMarkdown bestand via RStudio, en gebaseerd op jullie analyse de multiple choice vragen in te dienen.
- ▶ Meer details volgen volgende week tijdens de eerste officiële case study.

Vele slides bevatten R code dat aangegeven wordt door dit lettertype.

Deze kan soms moeilijk zijn om te lezen, maar zal steeds uitgelegd worden tijdens de les. De code kan erg nuttig zijn tijdens het studeren...

Vragen?

- ▶ Tijdens / na de les
- ▶ jeroen.gilis@ugent.be

De normale verdeling

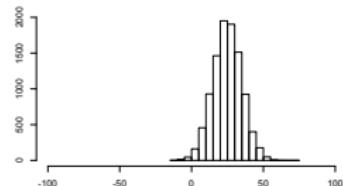
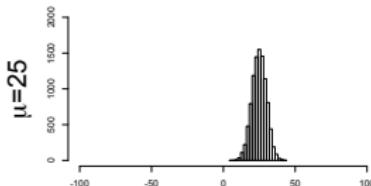
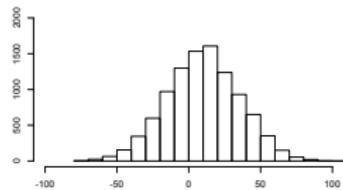
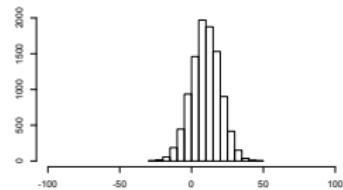
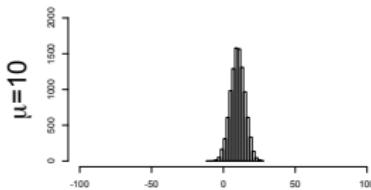
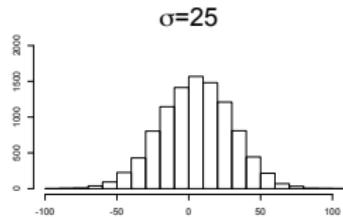
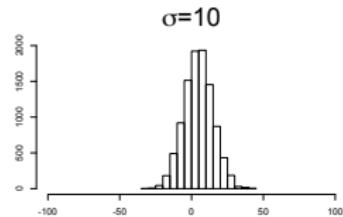
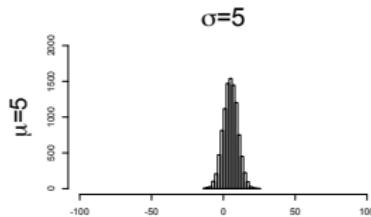
De normale verdeling is de meest gebruikte verdeling in de statistiek.

Het bevat enkele nuttige eigenschappen die erg belangrijk zijn in een statistische data-analyse.

Deze les zullen we vertrouwd raken met deze eigenschappen op een exploratieve wijze. In volgende lessen zullen we ze gebruiken om statistische tests uit te voeren.

De normale verdeling

De normale verdeling wordt beschreven door twee parameters: (populatie)gemiddelde μ en (populatie)variantie σ^2 (of standaarddeviatie σ). Deze bepalen respectievelijk de locatie en spreiding van de distributie.



De normale verdeling: interpretatie

Interpretatie van μ is eenvoudig: de centrale locatie van de distributie.

Interpretatie van σ is minder eenvoudig in kwantitatieve termen.

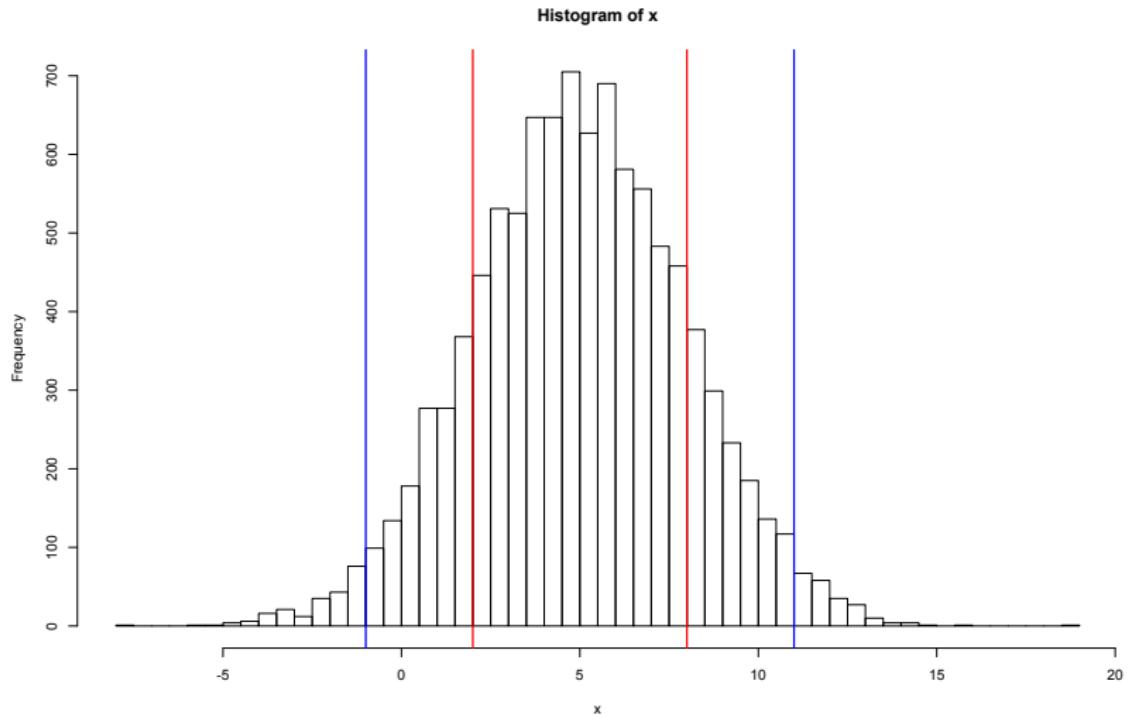
Volgende *rule of thumb* is echter toepasbaar op de normale distributie:

$[\mu - \sigma, \mu + \sigma]$ omslaat 68% van de data

$[\mu - 2\sigma, \mu + 2\sigma]$ omslaat 95% van de data

Dit is enkel geldig **indien de data een normale distributie volgen.**

De normale verdeling: interpretatie



De normale verdeling: interpretatie

In realiteit kennen we zelden het populatiegemiddelde μ of de populatievariantie σ^2 , waardoor we deze moeten benaderen door steekproef gemiddelde \bar{X} en steekproef variantie S^2 .

Dit induceert **sampling variabiliteit**:

```
> x <- rnorm(n=1e4, mean=5, sd=3)
> mean(x<5+3 & x>5-3)
[1] 0.6896
> mean(x<5+(2*3) & x>5-(2*3))
[1] 0.9576
```

De normale verdeling: interpretatie

In realiteit kennen we zelden het populatiegemiddelde μ of de populatievariantie σ^2 , waardoor we deze moeten benaderen door steekproef gemiddelde \bar{X} en steekproef variantie S^2 .

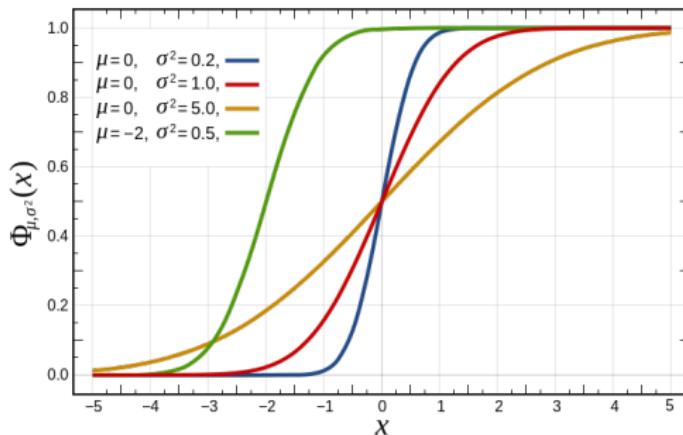
Dit induceert **sampling variabiliteit**:

```
> x <- rnorm(n=1e4, mean=5, sd=3)
> mean(x<5+3 & x>5-3)
[1] 0.6896
> mean(x<5+(2*3) & x>5-(2*3))
[1] 0.9576
```

→ De getallen komen niet exact overeen met de vermelde percentages vanwege sampling variabiliteit.

De cumulatieve distributiefunctie (CDF)

De cumulatieve distributiefunctie zegt welke fractie van de data kleiner is dan een bepaalde waarde x .



Deze fracties worden kwantielen of percentielen genoemd.

Percentielen/kwantielen van de normale verdeling

Een percentiel/kwantiel p kan men berekenen door de **cumulatieve distributiefunctie** (CDF) te evalueren met een waarde x .

$$\text{CDF: } P(X \leq x) = p,$$

i.e.: Wat is de kans p dat variabele X kleiner is dan of gelijk aan de waarde x ?

Indien $X \sim N(0, 1)$, dan $P(X \leq x) = \Phi(x)$.

Percentielen/kwantielen van de normale verdeling

Voorbeeld:

Stel: $X \sim N(5, 3^2)$

Bereken het percentiel/kwantiel voor $x = 5$: $P(X \leq 5) = ??$

¹ Gezien de symmetrie van de normale distributie is diens gemiddelde \bar{x} gelijk aan diens mediaan en de mediaan is het 50% percentiel/kwantiel van een distributie.

Percentielen/kwantielen van de normale verdeling

Voorbeeld:

Stel: $X \sim N(5, 3^2)$

Bereken het percentiel/kwantiel voor $x = 5$: $P(X \leq 5) = ??$

- ▶ Theoretisch verwachten we $P(X \leq 5) = 0.5$. Waarom? ¹

¹ Gezien de symmetrie van de normale distributie is diens gemiddelde \bar{x} gelijk aan diens mediaan en de mediaan is het 50% percentiel/kwantiel van een distributie.

Percentielen/kwantielen van de normale verdeling

Voorbeeld:

Stel: $X \sim N(5, 3^2)$

Bereken het percentiel/kwantiel voor $x = 5$: $P(X \leq 5) = ??$

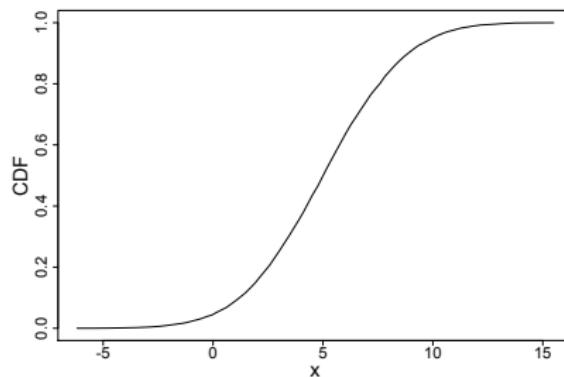
- ▶ Theoretisch verwachten we $P(X \leq 5) = 0.5$. Waarom? ¹
- ▶ Empirisch: zie volgende slide

¹ Gezien de symmetrie van de normale distributie is diens gemiddelde \bar{x} gelijk aan diens mediaan en de mediaan is het 50% percentiel/kwantiel van een distributie.

Percentielen/kwantielen van de normale verdeling

De CDF kan echter ook empirisch geschat worden obv de gesimuleerde data x :

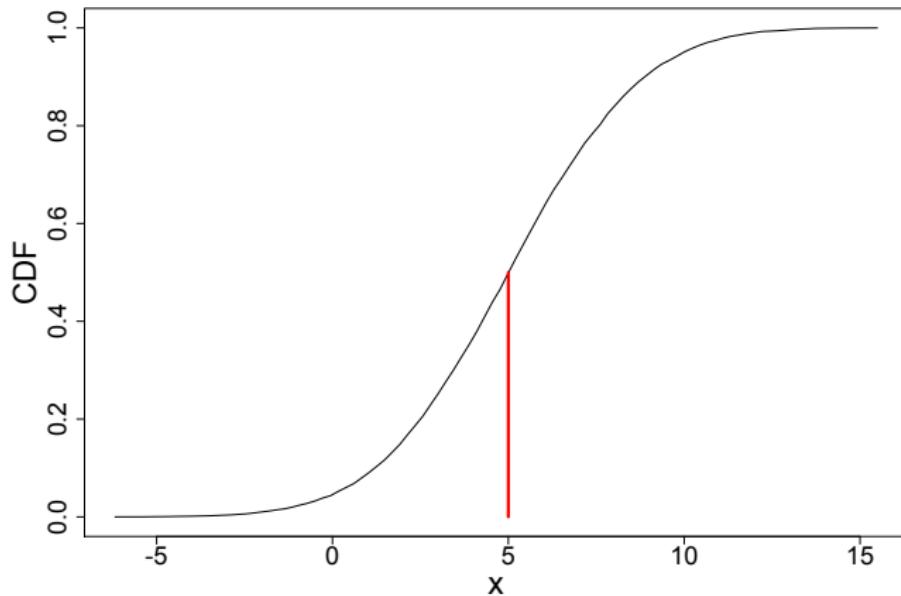
```
grid <- seq(min(x),max(x),length.out=100)
cdf <- vector(length=100)
for(i in 1:100){ cdf[i] <- mean(x<=grid[i]) }
plot(x=grid,y=cdf, type="l", ylab="CDF")
```



Dit zal echter niet gebruikt worden in deze cursus, maar de R code kan intuïtie geven in hoe men de CDF kan interpreteren.

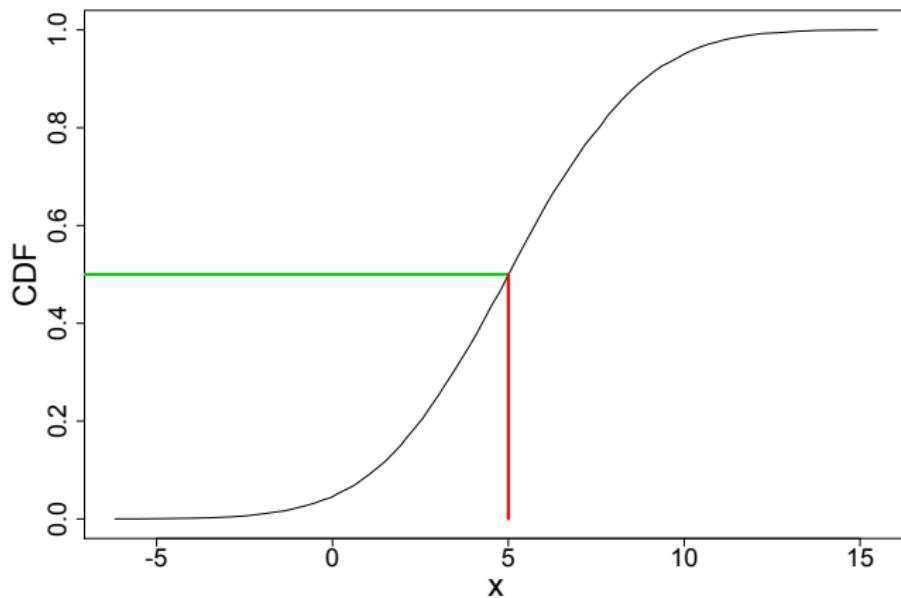
Percentielen/kwantielen van de normale verdeling

Bereken het percentiel/kwantiel voor $x = 5$ obv de empirische CDF:



Percentielen/kwantielen van de normale verdeling

Bereken het percentiel/kwantiel voor $x = 5$ obv de empirische CDF:



Percentielen/kwantielen van de normale verdeling

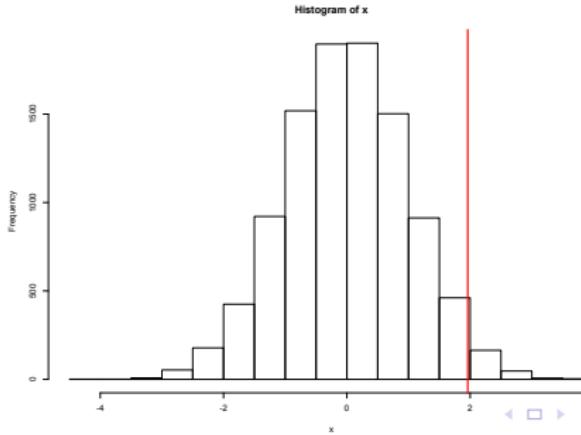
Bereken het percentiel/kwantiel voor $x = 5$ obv de empirische CDF:

```
mean(x<=5)  
[1] 0.4993
```

Opnieuw niet identiek gelijk aan de theoretische waarde vanwege sampling variabiliteit, maar ligt er erg dicht bij.

```
> x=rnorm(n=1e4, mean=0, sd=1) #standard normal
> mean(x<=1.96) #empirisch
[1] 0.9744
> pnorm(q=1.96, mean=0, sd=1) #theoretisch
[1] 0.9750021
> hist(x)
> abline(v=1.96, col="red", lwd=3)
```

Theoretisch ligt 97.5% van de data van een standaard normale $N(0, 1)$ distributie links van 1.96. Hoeveel ligt er dan links van -1.96?



Percentielen/kwantielen van de normale verdeling

Omgekeerd lukt ook: We kunnen de inverse van de CDF evalueren met een percentiel of kwantiel om de corresponderende waarde te verkrijgen.

Dit zullen we typisch doen om kritische waarden te berekenen bij statistische testen.

In R:

```
> qnorm(0.5,mean=5,sd=3)
[1] 5
```

Werken met de normale distributie via R

In onderstaande: $X \sim N(0, 1)$ en x is een willekeurige waarde uit X .

- ▶ `dnorm`: Wat is $P(X = x)$? ²
- ▶ `pnorm`: De CDF. Wat is $P(X \leq x)$ voor gegeven kwantiel x ?
- ▶ `qnorm`: Inverse CDF: Gegeven percentiel p , Voor welke x geldt $P(X \leq x) = p$?
- ▶ `rnorm`: Simuleer willekeurig observaties komende uit een normale verdeling.

Zie help files voor de vereiste argumenten! Bvb. `?qnorm`

²Eigenlijk, wat is $P(X > x - \delta \text{ \& } X < x + \delta)$. Uitleg volgt... 

Percentielen voor eender welke normale verdeling

- ▶ De R functies werken standaard met een $N(0, 1)$ verdeling.
- ▶ In de realiteit werken we vaak met normale verdelingen met een ander gemiddelde en/of variantie gebaseerd op de steekproef.
- ▶ We kunnen echter een normaal verdeelde geobserveerde variabele $X \sim N(\bar{X}, S_X^2)$ steeds transformeren naar een standaard normale verdeling $Z \sim N(0, 1)$

$$Z = \frac{X - \bar{X}}{S_X}$$

Dit zullen we doen in de volgende oefeningen. In R kan je dit doen via de argumenten `mean` en `sd` voor de functies op vorige slide.

Werken met de normale verdeling in R

Voor $X \sim N(2, 1)$. Bereken

- ▶ $P(X \geq 0)$
- ▶ q waarbij $P(X \leq q) = .8$

Werken met de normale verdeling in R

Voor $X \sim N(2, 1)$. Bereken

- ▶ $P(X \geq 0)$

```
> pnorm(q=0,mean=2,sd=1)
```

```
[1] 0.02275013
```

- ▶ q waarbij $P(X \leq q) = .8$

```
> qnorm(p=.8,mean=2,sd=1)
```

```
[1] 2.841621
```

Oefening 1

Uit een test die bij een groep studenten uit het 2de jaar Bachelor in de Biologie afgenoem werd, konden de volgende fictieve gegevens gehaald worden:

- ▶ Het gemiddelde gewicht bij de vrouwen is 57.22 kg met een SD van 6.39 kg. De gemiddelde lengte bij de vrouwen is 169 cm met een SD van 6.63 cm.
- ▶ Het gemiddelde gewicht bij de mannen is 68.73 kg met een SD van 8.36 kg. De gemiddelde lengte bij de mannen is 180.24 cm met een SD van 6.58 cm.

Als we veronderstellen dat zowel het gewicht als de lengte bij zowel de vrouwen als de mannen Normaal verdeeld is, bereken dan:

1. Welk percentage van de vrouwen zijn tussen 157 cm en 173 cm groot?
2. Welk percentage van de mannen wegen tussen 65 kg en 70 kg?

Oefening 2

Beschouw onderstaande tabel met betrekking tot de maximum dagelijkse pollutieniveaus per stad in 1986 tot 1989.

Tabel: Maximum Dairy Pollutant Levels by City, 1986 through 1989.
Pollutant level given as Mean \pm SD

| City | Carbon Monoxide | Nitrogen Dioxide | Sulfur Dioxide |
|-------------|-------------------|-------------------|-------------------|
| Los Angeles | 4.206 \pm 2.640 | 0.070 \pm 0.028 | 0.010 \pm 0.005 |
| Chicago | 2.510 \pm 1.002 | 0.045 \pm 0.013 | 0.025 \pm 0.011 |
| Milwaukee | 1.794 \pm 0.984 | 0.040 \pm 0.014 | 0.017 \pm 0.013 |

Oefening 2

1. Als deze gegevens bij benadering Normaal verdeeld zouden zijn, welk percentage van de observaties verwacht men dan binnen de aangegeven grenzen?
2. Stel dat 9.0 een kritisch niveau is voor koolstofmonoxide (Carbon Monoxide). Onder de assumptie van normaliteit, met welke kans wordt het kritisch niveau overschreden in L.A.?
3. Met welk constant getal moet men alle metingen daar verlagen om hoogstens 1% kans te hebben op overschrijding van de 9.0 limiet?
4. Observeer de gegevens voor “Sulfur Dioxide” in Milwaukee. Waarom kan de assumptie van normaliteit in dit geval niet voldaan zijn?

Hoe 'herken' je een normale distributie?

De meest gebruikte tool hiervoor is de QQ-plot.

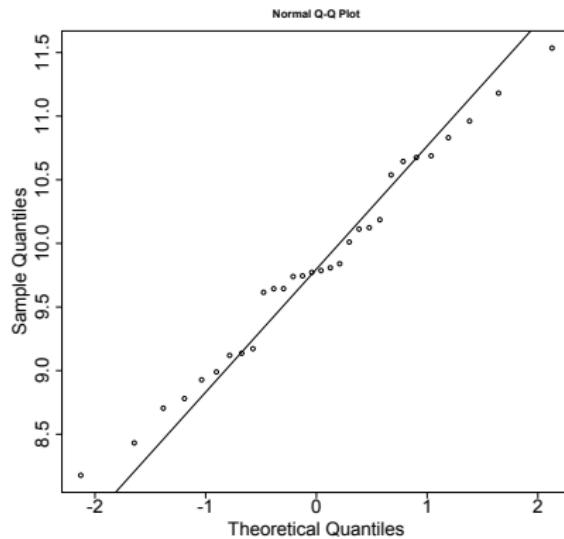
Constructie:

1. Datapunten van variabele X omzetten in percentielen
2. Deze percentielen gebruiken om corresponderende kwantielën van een $N(0, 1)$ distributie te berekenen
3. Zet datapunten X uit t.o.v. corresponderende $N(0, 1)$ waarden.

Hoe 'herken' je een normale distributie?

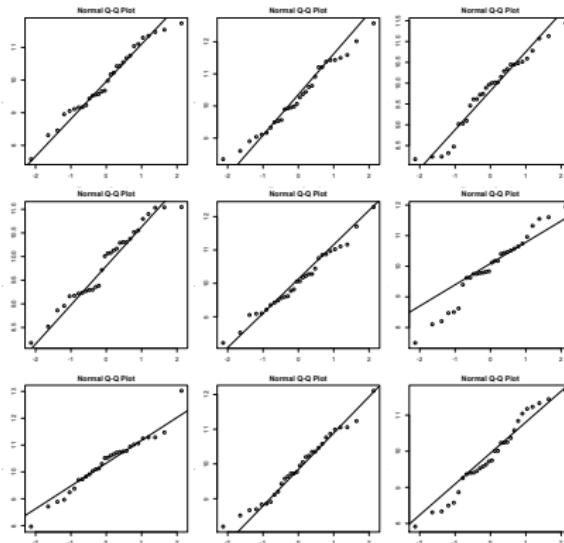
De QQ-plot vergelijkt de datapunten van de empirische steekproef met de corresponderende kwantilen van een normale verdeling.

Indien de kwantilen op een rechte lijn liggen hebben de distributies dezelfde vorm en kan men aannemen dat de steekproef een normale verdeling volgt.

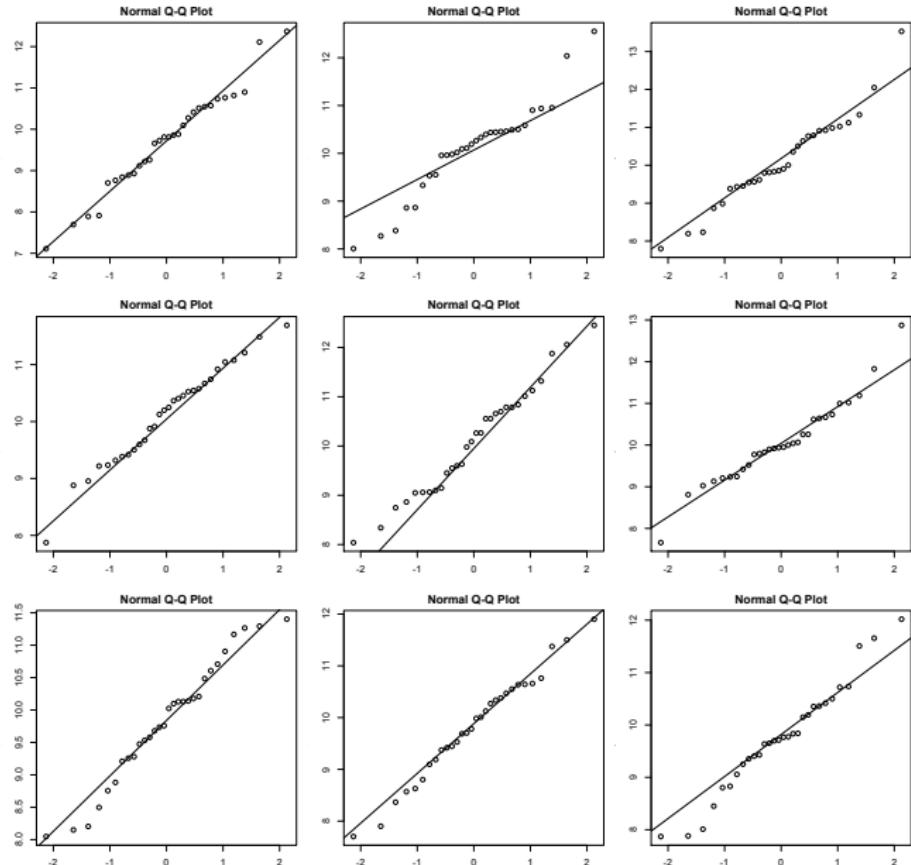


Evalueren van QQ-plots: sampling variabiliteit

```
par(mfrow=c(3,3),mar=c(3,3,2,1))
for(i in 1:9){
  x=rnorm(n=30, mean=10, sd=1) #sample data
  qqnorm(x) #qq plot
  qqline(x)}
```

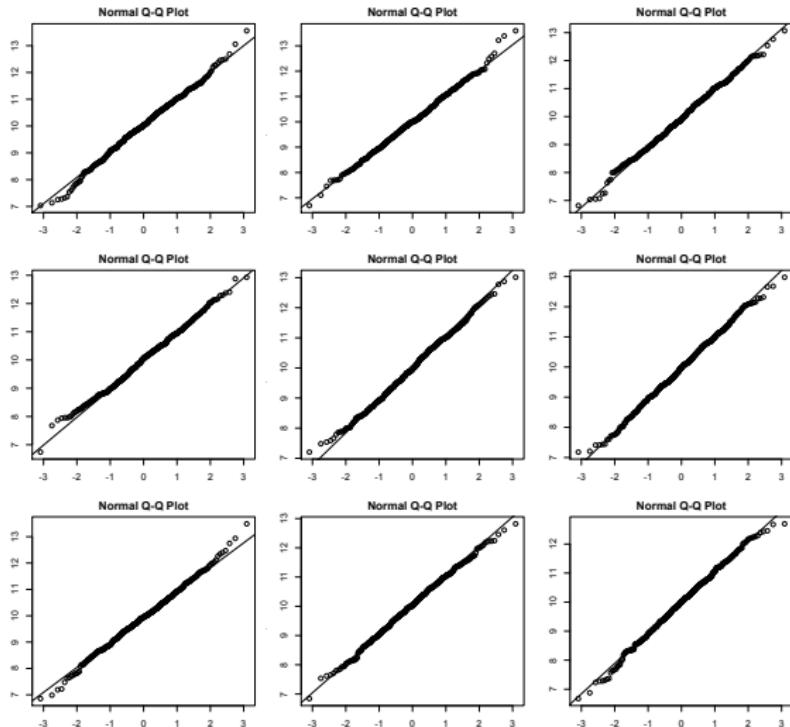


Evalueren van QQ-plots: sampling variabiliteit

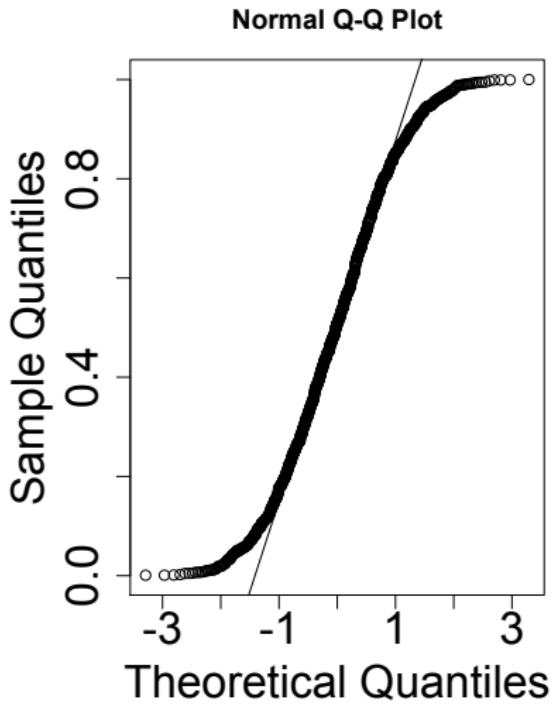
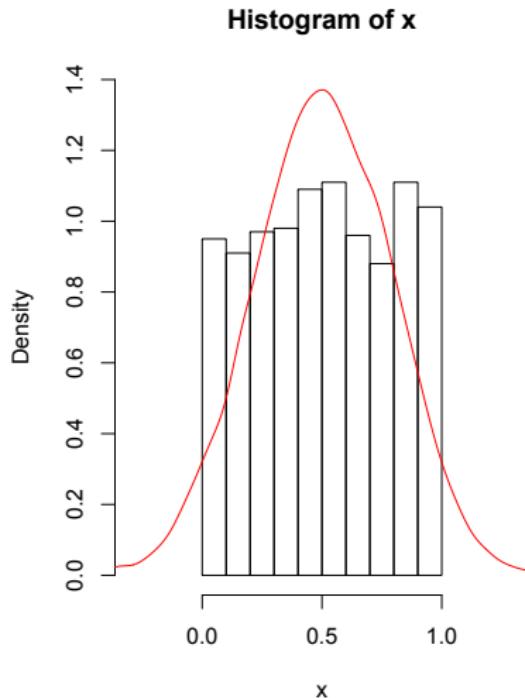


Evalueren van QQ-plots: sampling variabiliteit daalt met grotere steekproef

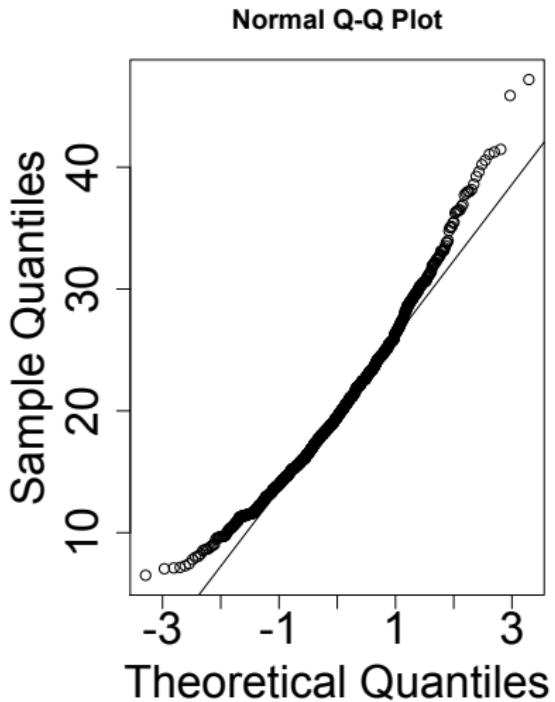
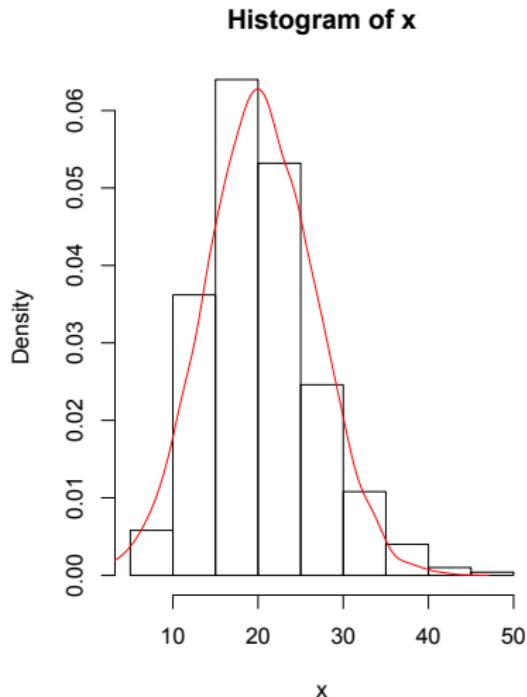
$n = 500$ ipv 30 op vorige slides



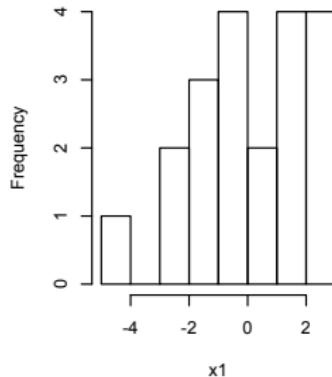
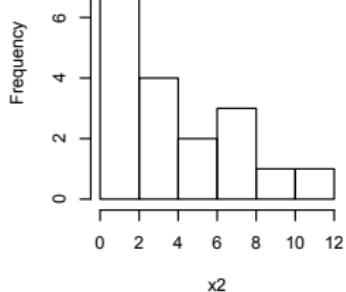
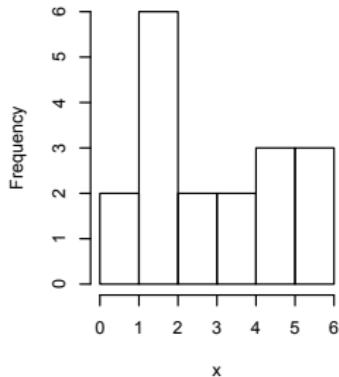
QQ-plot van uniforme verdeling (kort rechts en links)



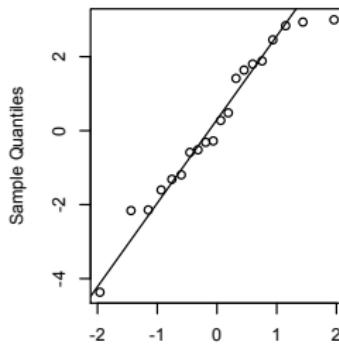
QQ-plot van χ^2 verdeling (kort links, scheef rechts)



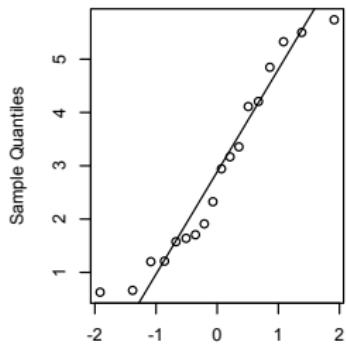
Test jezelf: welke QQ-plot hoort bij welk histogram?



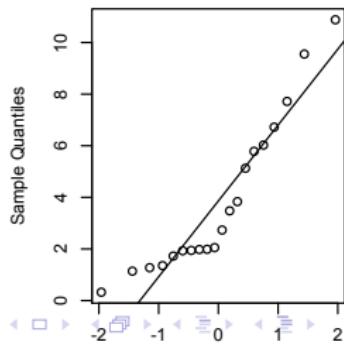
Normal Q-Q Plot



Normal Q-Q Plot



Normal Q-Q Plot



- ▶ Simuleer zelf QQ-plots
- ▶ Vul multiple choice vragen in op Minerva (opgelet, je hebt slechts één poging!)