HOWEST TOEGEPASTE INFORMATICA, 2022-2023, © BRIAN BAERT, DEVLIEGER INES

DATA ANALYTICS

HOOFDSTUK 5 - CONTINUE S.V. DE NORMALE VERDELING

howest.be

1

INLEIDING

- Tot nog toe hebben we enkel discrete stochastische variabelen bestudeerd = s.v.'s die een aftelbaar aantal waarden kunnen aannemen.
- In dit hoofdstuk: studie van **continue stochastische variabelen** = s.v.'s die een overaftelbaar aantal waarden kunnen aannemen.

  Symbolisch: $X: \Omega \to \mathbb{R}$

- Voorbeelden van experimenten die aanleiding geven tot continue s.v.'s: meten van het gewicht, lengte, IQ, … van personen.
- Bij continue s.v.'s heeft het geen zin om de kans op één bepaalde waarde te berekenen (waarom?) en berekent men eerder kansen dat X tot een bepaald interval behoort, zoals:

  $P(a \leq X \leq b)$ of $P(X \geq a)$ of $P(X \leq b)$.

2

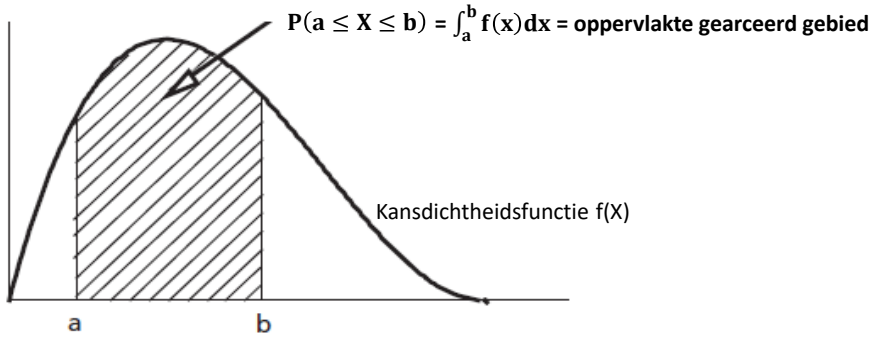
KANSDICHTHEIDSFUNCTIE

- Bij continue stochastische variabelen, worden kansen berekend aan de hand van hun **kansdichtheidsfunctie**. Dit is een functie die de verdeling van de kansen weergeeft.
- Een kansdichtheidsfunctie f van een continue s.v. moet aan onderstaande vereisten voldoen:
  1) f is een reële functie.
  2) Elke waarde f(x) is positief.
  3) $\int_{-\infty}^{+\infty} f(x)dx = 1$

3

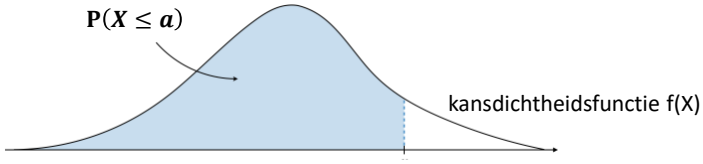
KANSBEREKENING BIJ EEN CONTINUE STOCHASTISCHE VARIABELE

De kans $\mathbf{P(a \leq X \leq b)}$is gelijk aan de integraal van a tot b van zijn kansdichtheidsfunctie, of anders gezegd: **de oppervlakte van het gebied dat begrensd wordt door de grafiek van die functie, de X-as en de rechten y=a en y=b.**



4

KANSBEREKENING BIJ EEN CONTINUE STOCHASTISCHE VARIABELE (2)



$P(X \geq a) = 1 - P(X < a)$

$P(-\infty \leq X \leq +\infty) = 1$

$P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$

5

GEMIDDELDE EN VARIANTIE VAN EEN CONTINUE STOCHASTISCHE VARIABELE

- Het gemiddelde of de verwachtingswaarde van een continue s.v. X wordt gedefinieerd door onderstaande formule:

$$\mu = E(X) = \int_{-\infty}^{+\infty} x f(x)\, dx$$

- De variantie van een continue s.v. X wordt gedefinieerd door onderstaande formule:

$$\text{Var(X)} = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$$

- Beide definities zijn een veralgemening van de definities van gemiddelde en variantie van een discrete s.v., waarbij het sigma-teken ($\sum$) (het symbool voor een aftelbare som) vervangen werd door het integraalteken ($\int$).

6

DE NORMALE VERDELING

- Veel metingen (IQ, BMI, lengte, gewicht,…) blijken heel vaak een hoeveelheid waarnemingsuitkomsten op te leveren die een zgn. **normale verdeling** benaderen.
- Definitie:
  De normale verdeling is een kansverdeling met als kansdichtheidsfunctie:

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

- Vermits deze functie afhankelijk is van de constanten μ en σ, vormen μ en σ de parameters van de normale verdeling.
- Notatie voor een normale verdeling X met parameters μ en σ: $\mathbf{X \sim N(\mu, \sigma)}$.

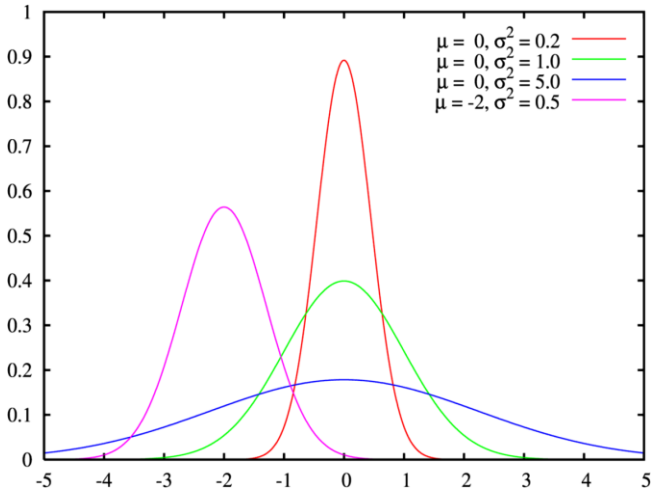
7

GRAFIEK VAN EEN NORMALE VERDELING

- De grafiek van een normale verdeling wordt een **Gausscurve** of **klokcurve** (Engels: **Bell curve**) genoemd.
- De ligging en de vorm van de klokcurve wordt bepaald door de paramaters μ en σ (zie volgende dia voor enkele voorbeelden).
- Enkele eigenschappen:
  → De klokcurve bereikt haar hoogtepunt als x = μ
  → De parameter σ geeft aan of de klokcurve spits of breed is: ze is spits bij een kleine waarde van σ en breed bij een grote waarde van σ
  → De klokcurve is symmetrisch met x= μ als symmetrieas
  → De x-waarden $\mu - \sigma$ en $\mu + \sigma$ bepalen de buigpunten van de grafiek

8

GRAFIEK VAN EEN NORMALE VERDELING = GAUSSCURVE OF KLOKCURVE



Figuur: Wikipedia

9

VERWACHTINGSWAARDE EN VARIANTIE VAN DE NORMALE VERDELING

- Eigenschap 1:
  De verwachtingswaarde van een normale verdeling $X \sim N(\mu, \sigma)$ is gelijk aan de parameter μ. Bovendien komt deze parameter ook overeen met de modus en de mediaan.

  In formulevorm: $\mathbf{E(x) = \mu}$

- Eigenschap 2:
  De standaardafwijking van een normale verdeling $X \sim N(\mu, \sigma)$ is gelijk aan de parameter σ.

  In formulevorm: $\boldsymbol{\sigma(X) = \sigma}$

10

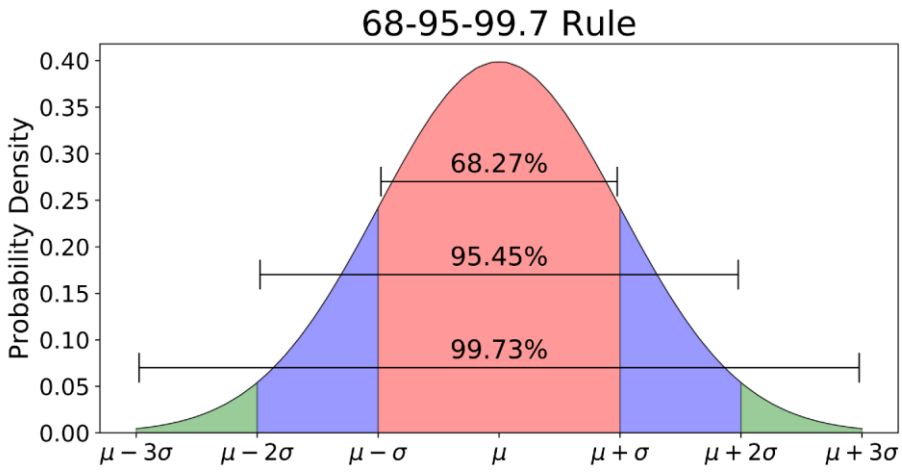
KANSEN BIJ EEN NORMALE VERDELING

- Als X~N, dan geldt:

$$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 68{,}27\%$$
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95{,}45\%$$
$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 99{,}73\%$$

- Onthoud:
  - Ongeveer **68%** van de waarnemingsgetallen van een normale verdeling wijken **hoogstens 1 keer de standaardafwijking af van het gemiddelde**.
  - Ongeveer 95% van de waarnemingsgetallen van een normale verdeling wijken hoogstens 2 keer de standaardafwijking af van het gemiddelde.
  - Ongeveer 99,7% van de waarnemingsgetallen van een normale verdeling wijken hoogstens 3 keer de standaardafwijking af van het gemiddelde.

GRAFIEK EN KANSEN VAN EEN NORMALE VERDELING



Figuur: http:// towardsdatascience.com

DE STANDAARDNORMALE VERDELING

- Definitie:
  De **standaardnormale verdeling** is een normale verdeling met **gemiddelde $\boldsymbol{\mu = 0}$** en **standaardafwijking $\boldsymbol{\sigma = 1}$**.
- Voor een standaardnormale verdeling gebruikt men meestal als symbool Z
  $$\Rightarrow \boldsymbol{Z \sim N(0, 1)}$$
- De kansdichtheidsfunctie van de standaardnormale verdeling Z is dus:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

13

STANDAARDISEREN

- Een normale verdeling X wordt omgevormd tot een standaardnormale verdeling Z met onderstaande formule:

$$\boldsymbol{Z = \frac{X - \mu}{\sigma}} \qquad \text{(z-score)}$$

- Deze bewerking wordt **standaardiseren** van een normaalverdeling genoemd.
- Het standaardiseren was vroeger (toen er gebruikgemaakt werd van tabellen) meestal de eerste stap bij het berekenen van kansen bij een normaalverdeling.
- De z-score wordt soms gebruikt om gegevens uit 2 verschillende normaalverdelingen met elkaar te kunnen vergelijken.

14

BEREKENING VAN KANSEN BIJ DE NORMALE VERDELING MBV EXCEL

- Als X normaal verdeeld is, dan kan een kans als P(X≤a) berekend worden mbv onderstaande functie in Excel:

$$P(X \leq a) = \text{NORM.VERD.N}(a;\mu;\sigma;\text{WAAR})$$

- Analoog kan, als Z standaardnormaal verdeeld is, een kans als P(Z≤a) berekend worden mbv onderstaande functie in Excel:

$$P(Z \leq a) = \text{NORM.S.VERD}(a;\text{WAAR})$$

- Het laatste argument van beide functies is een logische waarde die aanduidt of de cumulatieve verdelingsfunctie moet berekend worden (in geval van WAAR), of de kansdichtheidsfunctie (in geval van ONWAAR).

15

KANSBEREKENINGEN BIJ DE NORMALE VERDELING : VOORBEELD

Voorbeeld:

Bij een groep personen blijkt het IQ normaal verdeeld te zijn met een gemiddelde van 100 en een standaardafwijking van 15. Hoe groot is de kans dat het IQ van één van die personen tussen 85 en 115 ligt?

Bereken dit op 3 manieren:
- met gebruik van Excel-formule voor de normale verdeling
- met gebruik van Excel-formule voor de standaardnormale verdeling
- zonder gebruik van formules

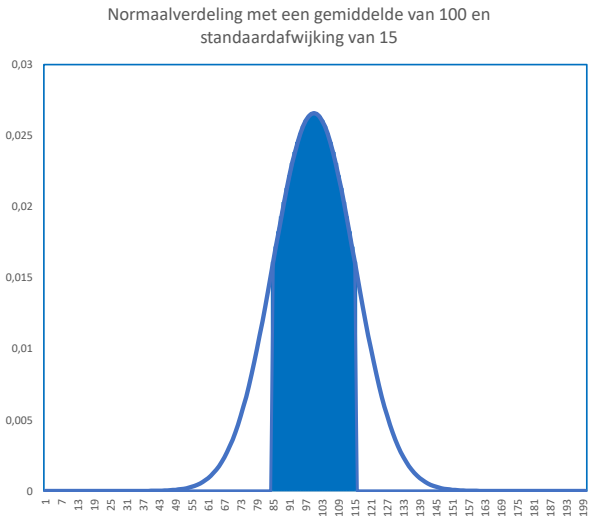
16

KANSBEREKENINGEN BIJ DE NORMALE VERDELING : VOORBEELD

Oplossing met Excel-formule voor de normale verdeling:

We moeten de oppervlakte berekenen van het blauwe gebied in de grafiek.

P(85 ≤ X ≤ 115) = P(X ≤ 115) – P(X ≤ 85)
= `NORM.VERD.N(115;100;15;waar)`
`-NORM.VERD.N(85;100;15;waar)`
≈ 0,84 – 0,16
≈ 0,68



17

KANSBEREKENINGEN BIJ DE NORMALE VERDELING : VOORBEELD - VERVOLG

- Oplossing met Excel-formule voor de standaardnormale verdeling:
  P(85 ≤ X ≤ 115) = P(X ≤ 115) – P(X ≤ 85)
  = P(Z ≤ $\frac{115-100}{15}$) - P(Z ≤ $\frac{85-100}{15}$) (transformeren naar een standaardnormaalverdeling = standaardiseren)
  = P (Z ≤ 1) – P(Z ≤ -1)
  = `NORM.S.VERD(1;WAAR) - NORM.S.VERD(-1;WAAR)`
  ≈ 0,84 – 0,16
  ≈ 0,68

- Oplossing zonder formules:
  We moeten nagaan hoe groot dat de kans is dat het IQ hoogstens 15 afwijkt v/h gemiddelde en omdat 15 in dit voorbeeld overeenkomt met de standaardafwijking, weten we eigenlijk zonder berekening al dat deze kans ongeveer 68% is.
  P(85 ≤ X ≤ 115) = P(100-15 ≤ X ≤ 100+15) = P(μ-σ ≤ X ≤ μ+σ) ≈ 68% (zie dia nr 12)

18

BEREKENING VAN WAARDEN ALS EEN KANS GEKEND IS

- Als $X \sim N(\mu, \sigma)$, dan kan een waarde a berekend worden als de kans P(X≤a) gekend is, m.b.v. onderstaande functie in Excel:

  NORM.INV.N(P(X≤a);μ;σ)

- Voorbeeld:
  Neem terug de normale verdeling uit het voorbeeld van dia nr 16. Voor welke waarde van het IQ geldt dat 80% van de personen een IQ heeft dat kleiner is dan die waarde?

  Oplossing:
  We moeten dus a berekenen als gegeven is dat $P(X \leq a) = 0{,}8$.
  a = NORM.INV.N(0,8;100;15) ≈ 112,6 cm

19

OEFENINGEN

- Oefeningenreeks 5.3.3 p. 60

20