HOWEST TOEGEPASTE INFORMATICA, 2022-2023, © BRIAN BAERT

DATA ANALYTICS

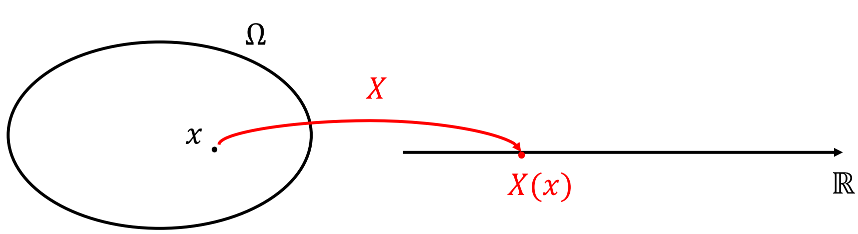
HOOFDSTUK 4 – DISCRETE STOCHASTISCHE VARIABELEN

howest.be

1

STOCHASTISCHE VARIABELEN

- Een **stochastische variabele** (afgekort als s.v.) bij een kansexperiment met uitkomstenruimte Ω, is een reële functie die aan elke mogelijke uitkomst x van dat experiment een **getal** toekent.
- Synoniemen: toevalsvariabele, kansvariabele (Engels: *random variable*).
- Meestal wordt een s.v. voorgesteld d.m.v. een hoofdletter X (of Y, Z, …).
- Dus $X: \Omega \rightarrow \mathbb{R}$.



2

SOORTEN STOCHASTISCHE VARIABELEN

- Men kan 2 belangrijke soorten stochastische variabelen onderscheiden:
  - **Discrete stochastische variabelen**
  - **Continue stochastische variabelen**
- Een s.v. X is **discreet** als X slechts een **eindig** of een **aftelbaar** aantal waarden aan kan nemen. Dergelijke kansvariabelen komen in dit hoofdstuk aan bod. We onderscheiden aftelbaar eindig (vb. binomiaalverdeling) en aftelbaar oneindig (vb. Poissonverdeling).
- Een s.v. X is **continu** als X een **overaftelbaar** (oneindig) aantal waarden aan kan nemen. Dergelijke kansvariabelen komen in volgend hoofdstuk aan bod.

3

KANSFUNCTIE

- Bij een stochastische variabele X, zijn we meestal geïnteresseerd in de **kans(verdelings)functie f** die de kansen berekent op elke mogelijke uitkomst van de stochastische variabele.
- Dus, voor elke gebeurtenis x_i geldt dan

$$f(x_i) = P(X = x_i)$$

- Bij een discrete s.v. zal de grafiek van deze functie een verzameling van afzonderlijke (en dus niet met elkaar verbonden) punten zijn.

4

VOORBEELD VAN EEN DISCRETE STOCHASTISCHE VARIABELE

Voorbeeld:
Bij een experiment met een “eerlijke” dobbelsteen wordt het aantal gegooide ogen uitbetaald.

$\Rightarrow$ s.v. X kan gedefinieerd worden als het bedrag dat uitbetaald wordt

X: 1 ==> f(1) = P(X=1) = $\frac{1}{6}$ X: 4 ==> f(4) = P(X=4) = $\frac{1}{6}$

X: 2 ==> f(2) = P(X=2) = $\frac{1}{6}$ X: 5 ==> f(5) = P(X=5) = $\frac{1}{6}$

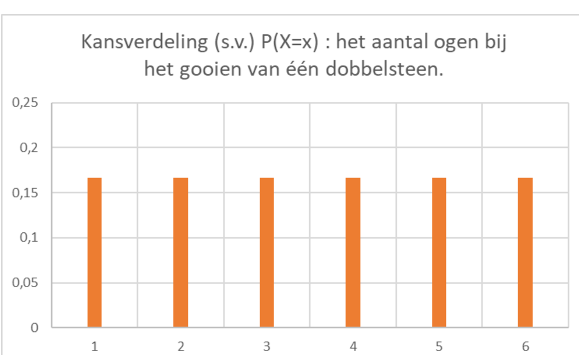
X: 3 ==> f(3) = P(X=3) = $\frac{1}{6}$ X: 6 ==> f(6) = P(X=6) = $\frac{1}{6}$

x	1	2	3	4	5	6
$P(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

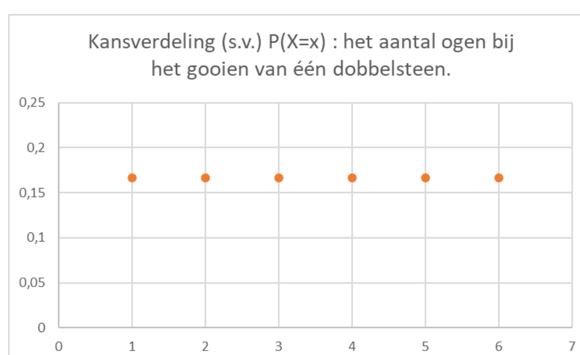
5

VOORBEELD V/D GRAFIEK V/E KANSVERDELING BIJ EEN DISCRETE STOCHASTISCHE VARIABELE

x	1	2	3	4	5	6
$P(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$



of



6

UNIFORME VERDELING

- In het vorige voorbeeld zijn de kansen op alle mogelijke waarden van de stochastische variabele even groot. Men spreekt in dit geval van een **uniforme of homogene verdeling**.
- Formele definitie:

  Een discrete s.v. X heeft een uniforme verdeling op n waarden $x_1, x_2, \dots, x_n$ als:

$$f(x_i) = P(X = x_i) = \frac{1}{n}$$

- De grafiek van een uniforme kansverdelingsfunctie bestaat uit een aantal punten die op gelijke hoogte van de horizontale as liggen (zie ook vorige dia).

7

VERWACHTINGSWAARDE VAN EEN DISCRETE S.V.

- Als we de kansen op de diverse waarden van een s.v. kennen, dan kunnen we op voorhand al een verwachting uitspreken over de waarde van een s.v. Men spreekt van de verwachtingswaarde, of de gemiddelde waarde van een stochastische variabele.
- **De verwachtingswaarde $\boldsymbol{E(X)}$** ("Expectancy", soms ook aangeduid met de Griekse letter $\boldsymbol{\mu}$) van een stochastische variabele wordt gedefinieerd als:

$$\boldsymbol{\mu = E(X) = \sum_{x_i \in \Omega} x_i . P(X = x_i) = \sum_{x_i \in \Omega} x_i . f(x_i)}$$

- De verwachtingswaarde is in feite een "gewogen" gemiddelde

8

VOORBEELD VAN VERWACHTINGSWAARDE VAN EEN DISCRETE S.V.

Voorbeeld 1:
Het verwachte aantal keer kop bij 3x opgooien van een munt wordt gegeven door de verdeling hieronder.
Bepaal op basis van de verdeling de verwachtingswaarde.

x	0	1	2	3
P(X=x)	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Oplossing:

$$E(X) = 0\,.\frac{1}{8} + 1\,.\frac{3}{8} + 2\,.\frac{3}{8} + 3\,.\frac{1}{8} = \frac{12}{8} = 1{,}5$$

Het verwachte aantal keer kop bij het 3x opgooien van een munt is 1,5 keer.

VOORBEELD VAN VERWACHTINGSWAARDE VAN EEN DISCRETE S.V.

Voorbeeld 2:
Hoeveel ogen zullen er gemiddeld gegooid worden bij een vervalste dobbelsteen waarbij er 5x zoveel kans is om een 6 te gooien als om een ander aantal ogen te gooien?

Oplossing:

x	1	2	3	4	5	6
P(X=x)	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{5}{10}$

$$E(X) = 1\,.\frac{1}{10} + 2\,.\frac{1}{10} + 3\,.\frac{1}{10} + 4\,.\frac{1}{10} + 5.\frac{1}{10} + 6.\frac{5}{10} = \frac{45}{10} = \frac{9}{2} = 4{,}5$$

VARIANTIE VAN EEN DISCRETE STOCHASTISCHE VARIABELE

- Een maat voor de **spreiding** van de verschillende waarden v/e stochastische variabele rond de gemiddelde waarde, is de **variantie**.
- **De variantie Var (X)** van een stochastische variabele X wordt gedefinieerd als de gemiddelde waarde van het kwadraat van de afwijkingen van de verschillende mogelijke waarden van X t.o.v. de gemiddelde waarde E(X).
- In formulevorm:

$$\sigma^2 = \text{Var}(X) \sum_{x_i \in \Omega} \left(X(x_i) - E(X)\right)^2 . P(X = x_i) = E((X - E(X)^2)$$

- Praktische formule voor het berekenen van de variantie:

$$\mathbf{Var}(\boldsymbol{X}) = \boldsymbol{E}\left(\boldsymbol{X}^{\mathbf{2}}\right) - \left(\boldsymbol{E}(\boldsymbol{X})\right)^{\mathbf{2}}$$

11

STANDAARDAFWIJKING VAN EEN DISCRETE STOCHASTISCHE VARIABELE

- In praktijk wordt er veel gebruikgemaakt van een andere, uit de variantie afgeleide spreidingsmaat, nl. de **standaardafwijking**.
- De **standaardafwijking** $\boldsymbol{\sigma}$ van een stochastische variabele wordt gedefinieerd als de vierkantswortel uit de variantie van die stochastische variabele.
- In formulevorm:

$$\sigma(X) = \sqrt{\text{Var}(X)}$$

12

VOORBEELD VAN VARIANTIE EN STANDAARDAFWIJKING VAN EEN DISCRETE S.V.

Voorbeeld:
Bereken de standaardafwijking van de s.v in dia 10 (voorbeeld 2).

Oplossing:
We berekenen de variantie mbv de formule: Var(X) = E(X²)-(E(X))²

$\boldsymbol{x_i}$	1	2	3	4	5	6
$\boldsymbol{P(X = x_i)}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{5}{10}$
$\boldsymbol{x_i^2 . P(X = x_i)}$	0,1	0,4	0,9	1,6	2,5	18

$E(X^2) = 0{,}1 + 0{,}4 + 0{,}9 + 1{,}6 + 2{,}5 + 18 = 23{,}5$
$\text{Var}(X) = E(X^2) - (E(X))2 = 23{,}5 - (4{,}5)^2 = 3{,}25$ (gemiddelde van X =4,5 zie dia 10)
$\sigma(X) = \sqrt{3{,}25} \approx 1{,}80$

DE BINOMIALE VERDELING

- Definitie:
  De **binomiale verdeling** is een discrete kansverdeling die een beschrijving geeft van het aantal successen dat kan optreden als een experiment een aantal keer herhaald wordt. Een dergelijk experiment noemt men een **Bernouilli-experiment**.

- Notatie:
  De s.v. X van een binomiale verdeling met n experimenten en met kans p op 1 succes, wordt soms genoteerd als: $\boldsymbol{X \sim B(n, p)}$.

- Voorbeelden:
  → Opgooien van een munt, kop of munt gooien
  → Onderdeel in een productieproces defect of niet (stat. Kwaliteitscontrole)
  → Kweken van reukerwten: rood of wit

DE FORMULE VAN DE BINOMIALE VERDELING

- Bij een binomiale verdeling $X \sim B(n, p)$ is de **kans op k successen**:

Aantal keer dat het experiment uitgevoerd wordt

kans op 1 succes

$$\boldsymbol{P(X = k) = C(n, k).\,p^{k}.\,q^{n-k}}$$

Aantal successen

kans op 1 mislukking => q=1-p

- We kunnen dus bij een binomiaal verdeelde stochastische variabele de kans op een aantal successen berekenen, als we het aantal herhaalde experimenten én de kans op 1 succes kennen.

VERWACHTING EN VARIANTIE BIJ DE BINOMIALE VERDELING

- Eigenschap:
  De verwachtingswaarde bij $X \sim B(n, p)$ wordt berekend met onderstaande formule:

$$\boldsymbol{E(X) = n.p}$$

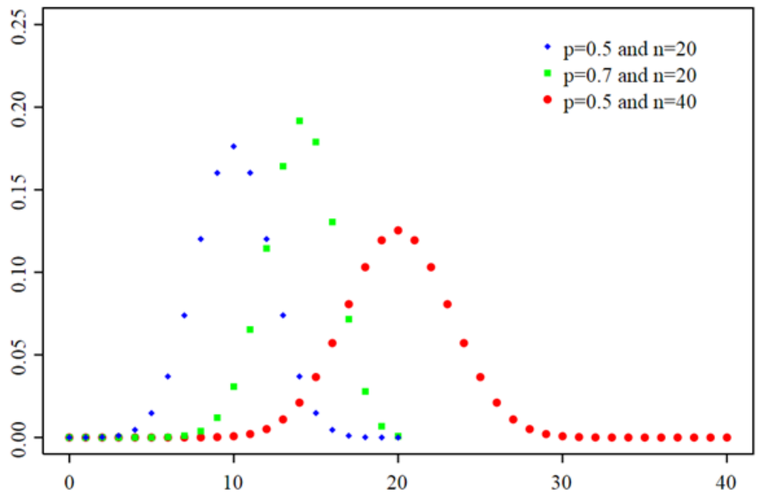
- Eigenschap:
  De variantie bij $X \sim B(n, p)$ wordt berekend met onderstaande formule:

$$\textbf{Var}(\boldsymbol{X}) = \boldsymbol{n.p.q}$$

- Eigenschap:
  De standaardafwijking bij $X \sim B(n, p)$ wordt berekend met onderstaande formule:

$$\boldsymbol{\sigma(X) = \sqrt{n.p.q}}$$

VOORBEELDEN VAN GRAFIEKEN VAN KANSVERDELINGSFUNCTIES BIJ DE BINOMIAALVERDELING



17

DE BINOMIALE VERDELING : VOORBEELDEN

<u>Voorbeeld 1</u>:
Iemand vult bij een multiple-choice-toets met 40 vierkeuzevragen de antwoorden willekeurig in.
Hoe groot is de kans dat hij 17 goede antwoorden heeft?

<u>Oplossing</u>:
Dit is een Binomiale verdeling met als parameters: $n = 40;\ p = \frac{1}{4} = 0{,}25$ en dus $X{\sim}B(40; 0{,}25)$

Vullen wij de formule in

$$P(X = k) = C(n, k).p^k.q^{n-k}$$

Dan krijgen wij

$$\mathrm{P(X = 17) = C(40{,}17)}.\left(\frac{1}{4}\right)^{17}.\left(\frac{3}{4}\right)^{23} \approx 0{,}0069$$

Of via Excel:

`P(X=17)=BINOM.VERD(17;40;0,25;ONWAAR)` $\approx 0{,}0069$

18

DE BINOMIALE VERDELING : VOORBEELDEN

Voorbeeld2:
Iemand vult bij een multiple-choice-toets met 20 vierkeuzevragen de antwoorden willekeurig in.
Bereken de kans dat hij ten hoogste 5 vragen goed heeft.

Oplossing:

$$X \sim B\left(20; \frac{1}{4}\right)$$

Wij moeten $P(X \leq 5)$ berekenen, dat is in feite:

$$P(X \leq 5) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$$

Deze 6 kansen kunnen elk apart berekend worden met de methode uitgelegd op de vorige dia, maar daar komt veel rekenwerk bij te pas.

Of via Excel:

```
P(X≤5)=BINOM.VERD(5;20;0,25;WAAR) ≈ 0,6172
```


DE POISSON-VERDELING

- Een ander bijzondere soort discrete kansverdeling is de **Poisson verdeling**. Deze verdeling is vaak bruikbaar voor het beschrijven van het aantal gebeurtenissen dat in een bepaalde tijdsperiode (of in een bepaald oppervlak of volume) zal optreden.
- Voorbeelden van s.v. waarvoor de Poisson verdeling een goed model is, zijn:
  - → Het aantal bedrijfsongevallen in een maand in een fabriek
  - → Het aantal negatieve reacties op een bepaald geneesmiddel.
  - → Het voorkomen van sterren in een bepaald deel van de ruimte.
- De Poisson verdeling hangt af van één parameter (λ genoteerd), wij noteren Po (λ)

$$\boldsymbol{X \sim Po(\lambda)}$$

- Formeel kan de Poisson verdeling kan benaderd worden door de Binomiale $Po(\lambda) \approx B(n,p)$ indien $n \to \infty, p \to 0$ en $np \to \lambda$

DE FORMULE VAN DE POISSON-VERDELING

- Als X een Poisson verdeling heeft met parameter λ, dan wordt de kans op k successen (of voorvallen, of registraties) binnen een bepaalde tijdsperiode, berekend met:

aantal successen (binnen een bepaalde tijdsperiode)

De verwachtingswaarde

$$\mathbf{P(X = k) = \frac{e^{-\lambda}.\lambda^{k}}{k!}}$$

getal van Euler: e=2,71828…

VERWACHTING EN VARIANTIE BIJ DE POISSON-VERDELING

- Eigenschap:
  De verwachtingswaarde bij $X \sim Po(\lambda)$ is gelijk aan de parameter λ.

$$\boldsymbol{E(X) = \lambda}$$

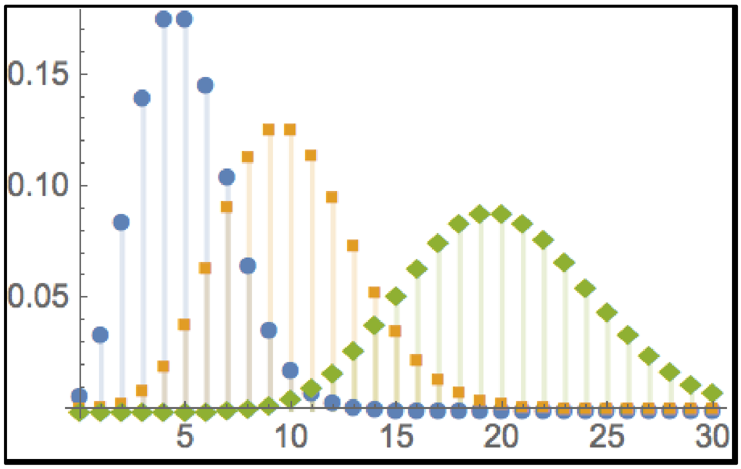
- Eigenschap:
  De variantie bij $X \sim Po(\lambda)$ is gelijk aan de parameter λ.

$$\mathbf{Var}(\boldsymbol{X}) = \boldsymbol{\lambda}$$

- Eigenschap:
  De standaardafwijking bij $X \sim Po(\lambda)$ wordt berekend met onderstaande formule:

$$\boldsymbol{\sigma(X) = \sqrt{\lambda}}$$

VOORBEELDEN VAN GRAFIEKEN VAN KANSVERDELINGSFUNCTIES BIJ DE POISSON-VERDELING



23

DE POISSON-VERDELING : VOORBEELD

Voorbeeld:
Op een helpdesk komen er gemiddeld 2 telefoongesprekken per minuut binnen.
a) Hoe groot is de kans dat er de volgende minuut 3 gesprekken binnenkomen?
b) Hoe groot is de kans dat er het volgend uur hoogstens 100 gesprekken binnenkomen?

Oplossing:
a) Stel X= "aantal gesprekken die per minuut binnenkomen" → $X \sim \text{Po}(2)$

$$\text{P}(\text{X} = 3) = \frac{e^{-2} . 2^3}{3!} \approx 0{,}18$$

Of via Excel:

P(X=3) = POISSON.VERD(3;2;ONWAAR) ≈ 0,18

b) Stel Y = "aantal gesprekken die per uur binnenkomen" → $Y \sim \text{Po}(120)$
Dit kan het snelst mbv onderstaande cumulatieve Poisson-verdelingsfunctie in Excel berekend worden:

P(X≤100) = POISSON.VERD(100;120;WAAR) ≈0,035

24

DE GEOMETRISCHE VERDELING

- Bij Bernoulli-experimenten kunnen we het aantal experimenten tellen totdat we een eerste keer succes hebben. Per definitie heeft de discrete s.v. X die we aldus bekomen, een **geometrische verdeling**. Dit wordt genoteerd als $X \sim \mathrm{Geom}(p)$, waarbij p de kans is op één succes.
- Het aantal mislukte Bernoulli-experimenten totdat we een eerste keer succes hebben, wordt berekend met de formule van de geometrische verdeling:

$$P(X = k) = (1 - p)^{k-1} . p$$

Kans op eerste keer succes bij het k-de deelexperiment

Kans op succes per deelexperiment

- Deze kans kan in Excel berekend worden m.b.v. de functie **`NEGBINOM.VERD(k-1;1;p;ONWAAR)`**

VERWACHTING EN VARIANTIE BIJ DE GEOMETRISCHE VERDELING

- Eigenschap:
De verwachtingswaarde bij $X \sim \mathrm{Geom}(p)$ wordt berekend met onderstaande formule:

$$\mathbf{E(X)} = \frac{\mathbf{1}}{\mathbf{p}}$$

- Eigenschap:
De variantie bij $X \sim \mathrm{Geom}(p)$ wordt berekend met onderstaande formule:

$$\mathbf{Var(X)} = \frac{\mathbf{1 - p}}{\mathbf{p^2}}$$

DE GEOMETRISCHE VERDELING : VOORBEELD

Voorbeeld:
Een dobbelsteen wordt een aantal maal na elkaar opgeworpen.
Hoe groot is de kans dat **je pas bij de 9de opgooi** een 6 gooit.

Oplossing:

We hebben hier te maken met een geometrische verdeling met parameter p = $\frac{1}{6}$.

Met de formule voor de geometrische verdeling:

$$P(X = 9) = \left(1 - \frac{1}{6}\right)^8 . \left(\frac{1}{6}\right) \approx 0{,}0387$$

Of via Excel:

P(X=9) = NEGBINOM.VERD(9−1;1;$\frac{1}{6}$;ONWAAR) $\approx 0{,}0387$

OEFENINGEN

- Oefeningenreeks 4.7 pagina 53.