

Univariate kansvariabelen

Sandra Van Aert

13 oktober 2011

Definitie

- ▶ kansvariabele
 - = stochastische variabele
 - = toevalsvariabele
- ▶ **definitie:** $X(\omega)$
functie die reëel getal associeert met elke uitkomst van een experiment
- ▶ **voorbeeld 1:** testen van een product
 - ▶ defect of niet-defect
 - ▶ $X(\text{defect}) = 0$ en $X(\text{niet-defect}) = 1$
- ▶ **voorbeeld 2:** nawegen van flessen in een vulproces
 - ▶ elke meting is reëel getal zodat $X(\omega) = \omega$

Notatie

- ▶ kansvariabele → hoofdletter

$X(\omega)$

X

Y, Z, X_1, X_2, \dots

= functies

- ▶ realisatie van kansvariabele → kleine letter

x, y, z

= getallen

- ▶ $P(X = x)$

$P(X < x)$

$P(x < 10)$ NIET

$P(X < 10)$ WEL

Voorbeeld

- ▶ experiment = opgooien 2 dobbelstenen
- ▶ 36 mogelijke uitkomsten ω_{ij}

(1,1) (1,2) ... (1,6)

(2,1) (2,2) ... (2,6)

\vdots \vdots \ddots \vdots

(6,1) (6,2) ... (6,6)

- ▶ mogelijke kansvariabelen

- ▶ X = som aantal ogen

$$X(\omega_{13}) = 4 = X(\omega_{22}) = X(\omega_{31})$$

- ▶ Y = absolute waarde verschil aantal ogen

$$Y(\omega_{13}) = 2 = Y(\omega_{31}) = Y(\omega_{24}) = \dots$$

Kansen

- ▶ experiment: gooien 2 dobbelstenen
- ▶ X = som aantal ogen
- ▶ $P(X = 7)$

$$= P((1, 6) \text{ gegooid of } (2, 5) \text{ gegooid of } (3, 4) \text{ gegooid of } (4, 3) \text{ gegooid of } (5, 2) \text{ gegooid of } (6, 1) \text{ gegooid})$$

$$= P((1, 6) \text{ gegooid}) + P((2, 5) \text{ gegooid}) + \dots + P((6, 1) \text{ gegooid})$$

$$= \frac{1}{36} + \frac{1}{36} + \dots + \frac{1}{36}$$

$$= \frac{6}{36} = \frac{1}{6}$$

Discrete kansvariabele

x	2	3	4	5	6	7	...	12
$P(X = x)$	1/36	2/36	3/36	4/36	5/36	6/36	...	1/36

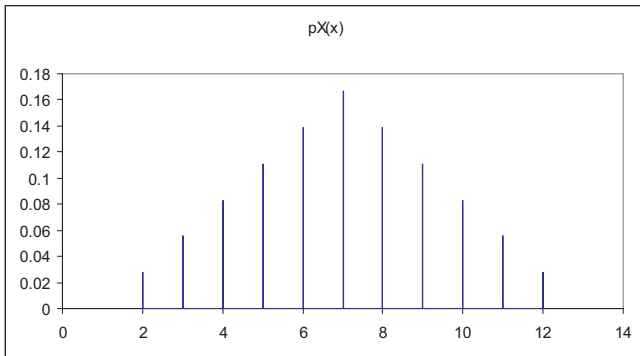
kansverdeling $p_X(x) = P(X = x)$

- ▶ $p_X(x) \geq 0$
- ▶ $\sum_{i=1}^k p_X(x_i) = 1$

merk op: soms $k \rightarrow \infty$

Kansverdeling grafisch

$$P(X = x)$$



(Cumulatieve) verdelingsfunctie

x	2	3	4	5	6	7	...	12
$P(X \leq x)$	1/36	3/36	6/36	10/36	15/36	21/36	...	1

cumulatieve verdelingsfunctie

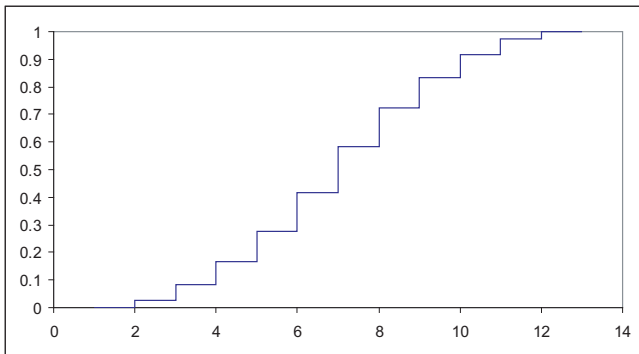
$$F_X(x) = P(X \leq x)$$

$$F_X(x) = \sum_{x_i \leq x} p_X(x_i), \quad \forall x$$

- ▶ niet dalend
- ▶ $F_X(-\infty) = 0$
- ▶ $F_X(+\infty) = 1$

Verdelingsfunctie grafisch

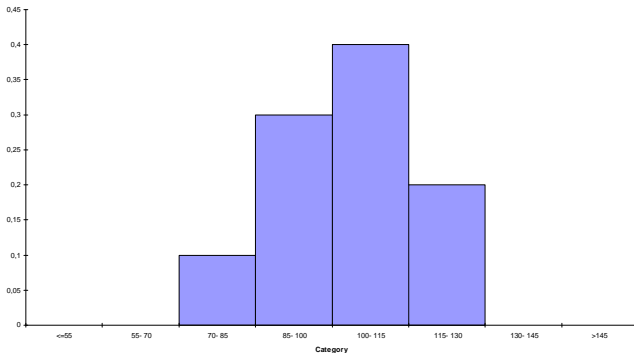
$$P(X \leq x)$$



Continue kansvariabele

Histogram met relatieve frequenties

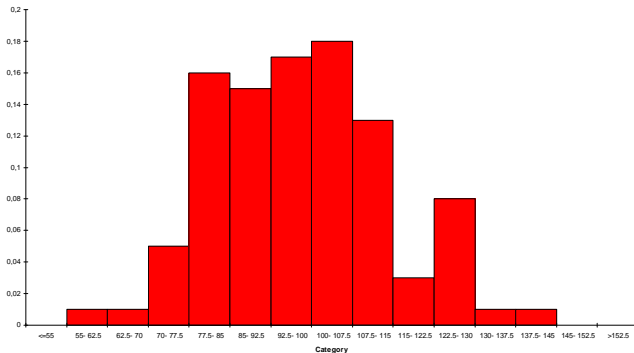
10 observaties



Continue kansvariabele

Histogram met relatieve frequenties

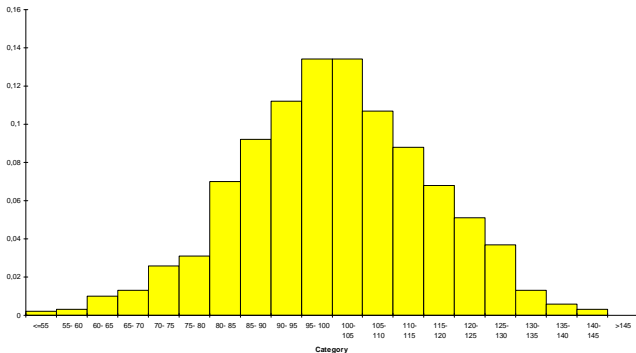
100 observaties



Continue kansvariabele

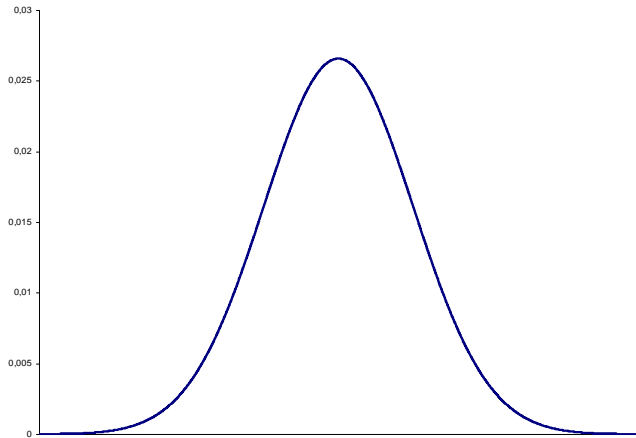
Histogram met relatieve frequenties

1000 observaties



Continue kansvariabele

Kansdichtheid



Kansdichtheid

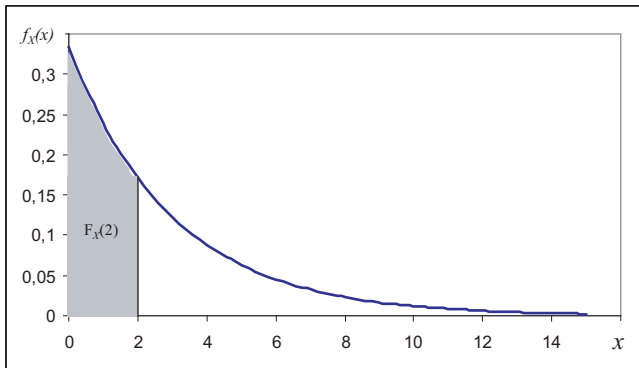
- ▶ a.h.w. polygoon
- ▶ functie $f_X(x)$

$$f_X(x) \geq 0$$

$$\int_{-\infty}^{+\infty} f_X(x) dx = 1$$

Kansdichtheid

- voorbeeld: $P(0 \leq X \leq 2)$



- ▶ kans = oppervlakte onder curve $f_X(x)$

$$P(a \leq X \leq b) = \int_a^b f_X(x) dx$$

- ▶ speciaal geval

$$P(X = a) = P(a \leq X \leq a) = \int_a^a f_X(x) dx = 0$$

Continu versus discreet

- ▶ kansdichtheid i.p.v. kansverdeling
- ▶ $f_X(x)$ i.p.v. $p_X(x)$
- ▶ $f_X(x) \geq 0 \leftrightarrow p_X(x) \geq 0$
- ▶ $\int_{-\infty}^{+\infty} f_X(x) dx = 1 \leftrightarrow \sum_{i=1}^k p_X(x_i) = 1$
- ▶ $f_X(x)$ soms $\geq 1 \leftrightarrow p_X(x)$ nooit > 1

Cumulatieve verdelingsfunctie

- ▶ voor elk reëel getal x

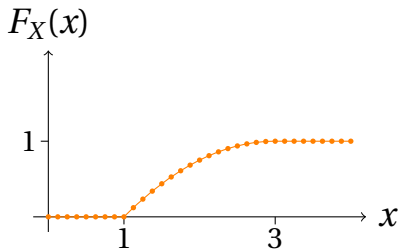
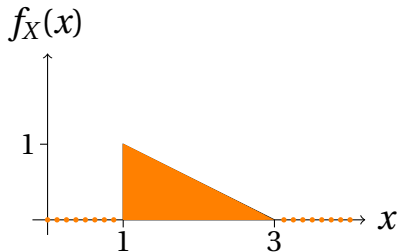
$$P(X \leq x)$$

$$= P(-\infty < X \leq x)$$

$$= \int_{-\infty}^x f_X(x) dx = \int_{-\infty}^x f_X(y) dy$$

$$= F_X(x) \quad (\text{cumulatieve verdelingsfunctie})$$

Cumulatieve verdelingsfunctie grafisch



Kansen en verdelingsfunctie

$$P(a \leq X \leq b) = \int_a^b f_X(x) dx$$

of

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

Opmerking

- ▶ $F_X(x)$
 $= P(-\infty < X \leq x)$
 $= \int_{-\infty}^x f_X(y) dy$
- ▶ $f_X(x)$ kan afgeleid worden uit $F_X(x)$
 $f_X(x)$
 $= \frac{d}{dx} F_X(x)$

Kengetallen

Sandra Van Aert

13 oktober 2011

Rekenkundig gemiddelde

- ▶ rekenkundig gemiddelde bij gegroeppeerde gegevens:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k f_i x_i = \frac{1}{n} (f_1 x_1 + f_2 x_2 + \cdots + f_n x_n)$$

- ▶ voorbeeld:

Aantal no-shows	0	1	2	3	4	5	6
Frequentie	11	38	32	9	6	3	1
Rel. frequentie	11%	38%	32%	9%	6%	3%	1%

$$\begin{aligned}\bar{x} &= \frac{1}{100} (11 \times 0 + 38 \times 1 + 32 \times 2 + 9 \times 3 + 6 \times 4 + 3 \times 5 + 1 \times 6) \\ &= 1.74\end{aligned}$$

Verwachte waarde of gemiddelde

$$\mu_X = E(X)$$

► discreet: $\sum_{i=1}^k x_i p_X(x_i)$

*analoog aan steekproefgemiddelde
gegroepeerde gegevens (hoofdstuk 3)*

► continu: $\int_{-\infty}^{+\infty} x f_X(x) dx$

Voorbeeld

- ▶ meerkeuzevragen
- ▶ 4 antwoordmogelijkheden
- ▶ slechts 1 van de 4 is juist
- ▶ puntenverdeling:
 - ▶ juist antwoord: +1
 - ▶ fout antwoord: $-1/3$
 - ▶ geen antwoord: 0
- ▶ heb je er belang bij te gokken als je geen enkel antwoord kunt uitsluiten?
- ▶ heb je er belang bij te gokken als je één antwoord kunt uitsluiten?

Voorbeeld

Kansverdeling bij gokken:

Antwoord	1	2	3	4
Kans	1/4	1/4	1/4	1/4

$$E(\text{punten}) = 1\frac{1}{4} - \frac{1}{3}\frac{1}{4} - \frac{1}{3}\frac{1}{4} - \frac{1}{3}\frac{1}{4} = 0$$

Voorbeeld

Kansverdeling bij 1 eliminatie:

Antwoord	1	2	3	4
Kans	1/3	1/3	1/3	0

$$E(\text{punten}) = 1\frac{1}{3} - \frac{1}{3}\frac{1}{3} - \frac{1}{3}\frac{1}{3} = \frac{1}{9}$$

Verwachte waarde van functie $Y = g(X)$

algemeen

- ▶ discreet: $\mu_Y = E(Y) = \sum_{i=1}^k g(x_i) p_X(x_i)$
- ▶ continu: $\mu_Y = E(Y) = \int_{-\infty}^{+\infty} g(x) f_X(x) dx$

lineaire functie $Y = aX + b$

- ▶ $\mu_Y = a\mu_X + b$
- ▶ $E(Y) = E(aX + b)$
$$= \int_{-\infty}^{+\infty} (ax + b) f_X(x) dx$$
$$= a \int_{-\infty}^{+\infty} x f_X(x) dx + b \int_{-\infty}^{+\infty} f_X(x) dx$$
$$= a\mu_X + b$$

Steekproefvariantie

- ▶ steekproefvariantie bij gegroepeerde gegevens:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^k f_i (x_i - \bar{x})^2$$

- ▶ voorbeeld:

Aantal no-shows	0	1	2	3	4	5	6
Frequentie	11	38	32	9	6	3	1
Rel. frequentie	11%	38%	32%	9%	6%	3%	1%

- ▶ $s^2 = \frac{1}{99} (11 \times (0 - 1.74)^2 + 38 \times (1 - 1.74)^2 + \dots + 1 \times (6 - 1.74)^2) = 1.53$

Variantie

σ_X^2 of $\text{var}(X)$

= verwachte waarde van $Y = g(X) = (X - \mu_X)^2$

► discreet:
$$\sigma_X^2 = \sum_{i=1}^k g(x_i) p_X(x_i)$$
$$= \sum_{i=1}^k (x_i - \mu_X)^2 p_X(x_i)$$

► Continu:
$$\sigma_X^2 = \int_{-\infty}^{+\infty} g(x) f_X(x) dx$$
$$= \int_{-\infty}^{+\infty} (x - \mu_X)^2 f_X(x) dx$$

$$\begin{aligned}\sigma_X^2 &= E(Y) = E[(X - \mu_X)^2] \\ &= E(X^2 - 2\mu_X X + \mu_X^2) \\ &= E(X^2) - 2\mu_X E(X) + \mu_X^2 \\ &= E(X^2) - 2\mu_X^2 + \mu_X^2 \\ &= E(X^2) - \mu_X^2\end{aligned}$$

Variantie lineaire functie $Y = aX + b$

$$\begin{aligned}\sigma_Y^2 &= E[(Y - \mu_Y)^2] \\&= E[(aX + b - a\mu_X - b)^2] \\&= E[(aX - a\mu_X)^2] \\&= E[a^2(X - \mu_X)^2] \\&= a^2 E[(X - \mu_X)^2] \\&= a^2 \sigma_X^2\end{aligned}$$

Nog meer begrippen

standaarddeviatie

$$\sigma_X = +\sqrt{\sigma_X^2}$$

gestandaardiseerde kansvariabele

$$Z = \frac{X - \mu_X}{\sigma_X}$$

$$E(Z) = 0 \text{ en } \sigma_Z^2 = \text{var}(Z) = 1$$

modus

$p_X(x)$ of $f_X(x)$ maximaal

Mediaan $\gamma_{0.5}$

- ▶ discreet

$$(1) \quad P(X \leq \gamma_{0.5}) \geq \frac{1}{2}$$

$$(2) \quad P(X \geq \gamma_{0.5}) \geq \frac{1}{2}$$

- ▶ continu

$$\frac{1}{2} = \int_{-\infty}^{\gamma_{0.5}} f_X(x) dx = \int_{\gamma_{0.5}}^{+\infty} f_X(x) dx$$

$(100 \times p)$ de percentiel γ_p

- ▶ continu

$$(1) \quad p = \int_{-\infty}^{\gamma_p} f_X(x) dx$$

$$(2) \quad 1 - p = \int_{\gamma_p}^{+\infty} f_X(x) dx$$

- ▶ $\gamma_{0.25}$ = eerste kwartiel
- ▶ $\gamma_{0.5}$ = tweede kwartiel = mediaan
- ▶ $\gamma_{0.75}$ = derde kwartiel

- ▶ Pearsons populatiescheefheidscoëfficiënt

$$SP^{pop} = \frac{3(\mu_X - \gamma_{0.5})}{\sigma_X}$$

- ▶ $-3 \leq S_p \leq +3$
- ▶ symmetrische verdeling : $SP^{pop} = 0$
- ▶ rechtsscheve verdeling : $SP^{pop} > 0$
- ▶ linksscheve verdeling : $SP^{pop} < 0$

- ▶ **scheefheidscoëfficiënt**

$$\text{scheefheidscoëfficiënt} = \frac{E[(X - \mu_X)^3]}{\sigma_X^3}$$

- ▶ symmetrische verdeling :
 $\text{scheefheidscoëfficiënt} = 0$
- ▶ rechtsscheve verdeling :
 $\text{scheefheidscoëfficiënt} > 0$
- ▶ linksscheve verdeling :
 $\text{scheefheidscoëfficiënt} < 0$

Discrete kansverdelingen

Sandra Van Aert

13 oktober 2011

Uniforme kansverdeling

alle mogelijke uitkomsten hebben evenveel kans

voorbeeld

- ▶ X = aantal ogen gegooid met 1 dobbelsteen
- ▶ 6 mogelijkheden
- ▶ $p_X(x) = \frac{1}{6}, \quad x = 1, 2, \dots, 6$

Algemeen

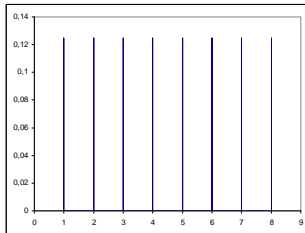
- ▶ stel X neemt gehele waarden aan $1, 2, \dots, N$
- ▶ kansverdeling:

$$p_X(n) = \frac{1}{N}, \quad n = 1, \dots, N$$

- ▶ cumulatieve verdelingsfunctie:

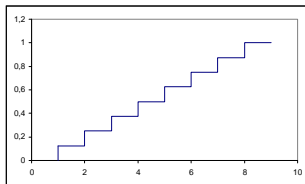
$$F_X(n) = P(X \leq n) = \sum_{i=1}^n \frac{1}{N} = \frac{n}{N}, \quad n = 1, \dots, N$$

$$P(X = n)$$



uniforme kansverdeling met $N = 8$

$$P(X \leq n)$$



cumulatieve verdelingsfunctie met $N = 8$

Bernoulli verdeling

kansvariabele kan waarde 0 of 1 aannemen

$X \rightarrow 1$ met kans π

$X \rightarrow 0$ met kans $1 - \pi$

$$p_X(x; \pi) = \pi^x (1 - \pi)^{1-x}, \quad x = 0, 1$$

π is parameter van Bernoullifamilie ($0 \leq \pi \leq 1$)

$$P(X = 1) = ?$$

$$= p_X(1; \pi)$$

$$= \pi^1 (1 - \pi)^{1-1}$$

$$= \pi (1 - \pi)^0$$

$$= \pi$$

$$P(X = 0) = ?$$

$$= p_X(0; \pi)$$

$$= \pi^0 (1 - \pi)^{1-0}$$

$$= 1 - \pi$$

$$\begin{aligned}\mu_X = E(X) &= \sum_i x_i p_X(x_i; \pi) \\ &= 1 \cdot p_X(1; \pi) + 0 \cdot p_X(0; \pi) \\ &= 1 \cdot \pi + 0 \cdot (1 - \pi) \\ &= \pi\end{aligned}$$

$$\begin{aligned}\sigma_X^2 = \text{var}(X) &= \sum_i (x_i - \mu_X)^2 p_X(x_i; \pi) \\ &= (1 - \pi)^2 \cdot p_X(1; \pi) + (0 - \pi)^2 \cdot p_X(0; \pi) \\ &= (1 - \pi)^2 \cdot \pi + \pi^2 \cdot (1 - \pi) \\ &= \pi(1 - \pi) [(1 - \pi) + \pi] \\ &= \pi(1 - \pi)\end{aligned}$$

Voorbeeld

- ▶ $X \rightarrow 1$ indien defect
 $X \rightarrow 0$ indien niet-defect
- ▶ $X \rightarrow 1$ indien geslaagd
 $X \rightarrow 0$ indien niet-geslaagd
- ▶ $X \rightarrow 1$ rode reukerwt
 $X \rightarrow 0$ witte reukerwt
- ▶ **Bernoulli proces/experiment:** *kansproces waarbij één element uit een Bernoulli verdeling gegenereerd wordt; successen (“1”) en falingen (“0”)*

Binomiale verdeling

- ▶ n opeenvolgende Bernoulli experimenten
- ▶ tel aantal successen X
- ▶ $X: 0, 1, 2, \dots, n$
- ▶ alle Bernoulli experimenten zijn onafhankelijk en hebben parameter π

Afleiding kansverdeling

- ▶ wat is kans op 3 successen en 5 falingen bij $n = 8$?
- ▶ stel:

$$\begin{array}{cccccccc} S, & F, & S, & F, & F, & F, & S, & F, \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ \pi & 1 - \pi & \pi & 1 - \pi & 1 - \pi & 1 - \pi & \pi & 1 - \pi \end{array}$$

- ▶ vermenigvuldigingsregel:
kans op deze volgorde =

$$\pi(1 - \pi)\pi(1 - \pi)(1 - \pi)(1 - \pi)\pi(1 - \pi) = \pi^3(1 - \pi)^5$$

Vervolg afleiding kansverdeling

- ▶ er zijn nog andere volgordes met 3 successen:

$$\frac{8!}{3!5!} \text{ volgordes met 3 successen}$$

- ▶ wegens de optelregel is de kans op 3 successen (en 5 falingen) bij $n = 8$ dan

$$\frac{8!}{3!5!} \pi^3 (1 - \pi)^5 = \binom{8}{3} \pi^3 (1 - \pi)^5$$

Definitie

- ▶ kans op x successen:

$$\begin{aligned}P(X = x) &= p_X(x; \pi, n) \\&= \frac{n!}{x! (n-x)!} \pi^x (1-\pi)^{n-x} \\&= \binom{n}{x} \pi^x (1-\pi)^{n-x}\end{aligned}$$

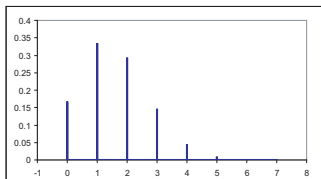
- ▶ parameters: n en π
- ▶ kengetallen:

$$\begin{aligned}\mu_X &= n\pi \\ \sigma_X^2 &= n\pi(1-\pi)\end{aligned}$$

$$\begin{aligned}\mu_X &= E(X) = E(Y_1 + Y_2 + \cdots + Y_n) \\ &= E(Y_1) + E(Y_2) + \cdots + E(Y_n) \\ &= \pi + \pi + \cdots + \pi \\ &= n\pi\end{aligned}$$

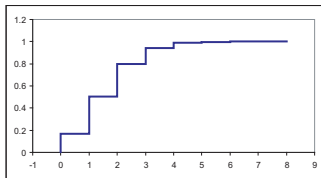
$$\begin{aligned}\sigma_X^2 &= \text{var}(X) = \text{var}(Y_1 + Y_2 + \cdots + Y_n) \\ &= \text{var}(Y_1) + \text{var}(Y_2) + \cdots + \text{var}(Y_n) \\ &= \pi(1 - \pi) + \pi(1 - \pi) + \cdots + \pi(1 - \pi) \\ &= n\pi(1 - \pi)\end{aligned}$$

$$P(X = x)$$



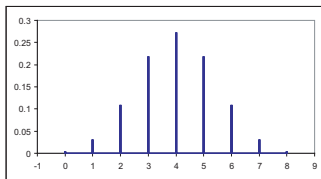
kansverdeling met $n = 8$ en $\pi = 0.2$

$$P(X \leq x)$$



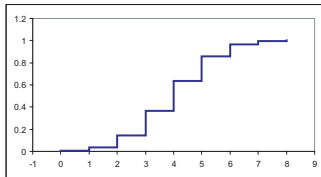
cumulatieve verdelingsfunctie met $n = 8$ en $\pi = 0.2$

$$P(X = x)$$



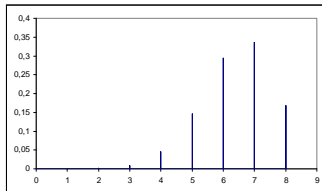
kansverdeling met $n = 8$ en $\pi = 0.5$

$$P(X \leq x)$$



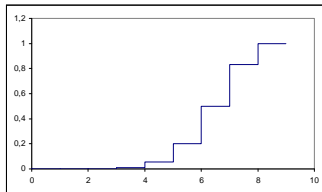
cumulatieve verdelingsfunctie met $n = 8$ en $\pi = 0.5$

$$P(X = x)$$



kansverdeling met $n = 8$ en $\pi = 0.8$

$$P(X \leq x)$$



cumulatieve verdelingsfunctie met $n = 8$ en $\pi = 0.8$

Voorbeeld

- ▶ $\pi = 0.10$ defectenratio
- ▶ $n = 20$ inspecties
- ▶ $P(X = 2) = ?$

- ▶ R: `dbinom(2, 20, 0.1)`
Matlab: `binopdf(2, 20, 0.1)`
 - ▶ de waarde van x
 - ▶ de parameter n
 - ▶ de parameter π

Vervolg voorbeeld

- ▶ $\pi = 0.10$ defectenratio
- ▶ $n = 20$ inspecties
- ▶ $P(X = 2) = ?$
- ▶ rekentoestel:

$$P(X = 2) = \frac{20!}{18!2!} (0.1)^2 (0.9)^{18} = 0.2852$$

Vervolg voorbeeld

- ▶ $P(X \geq 3) = ?$
- ▶ $P(X \geq 3) = 1 - P(X < 3) = 1 - P(X \leq 2)$
- ▶ R: `"=1-pbinom(2,20,0.1)`
Matlab: `"=1-binocdf(2,20,0.1)`
- ▶ rekentoestel

$$\begin{aligned} P(X \geq 3) &= 1 - P(X \leq 2) \\ &= 1 - P(X = 0) - P(X = 1) - P(X = 2) \\ &= \dots \end{aligned}$$

voorbeelden

- ▶ aantal defecten / lengte-eenheid
- ▶ aantal aardbevingen / tijdseenheid
- ▶ aantal bacteriën / volume-eenheid

voorwaarden Poisson proces

- ▶ gebeurtenissen komen niet in groep voor
- ▶ kans op gebeurtenis constant
- ▶ onafhankelijk in twee niet-overlappende intervallen

Poissonverdeling

- ▶ X : aantal keer dat een gebeurtenis voorkomt in een gegeven tijdsinterval
- ▶ tijdsinterval opsplitsen in zeer groot aantal (n) kleine deelintervallen
- ▶ n opeenvolgende Bernouilli experimenten
- ▶ X binomiaal verdeeld
- ▶ Poisson verdeling is limiet van de binomiaal verdeling $n \rightarrow \infty, \pi \rightarrow 0$ en $n\pi \rightarrow \lambda$

$$Poisson(\lambda) \approx bin(n, \pi)$$

Vergelijking Poisson en binomiale verdeling

	binomiale verdeling				Poisson
n	5	20	100	500	$\lambda = 1$
π	0.2	0.05	0.01	0.002	
$P(X = 0)$	0.3277	0.3585	0.3660	0.3675	0.3679
$P(X = 1)$	0.4096	0.3774	0.3697	0.3682	0.3679
$P(X = 2)$	0.2048	0.1887	0.1849	0.1841	0.1839
$P(X = 3)$	0.0512	0.0596	0.0610	0.0613	0.0613
$P(X = 4)$	0.0064	0.0133	0.0149	0.0153	0.0153

Definitie

$$p_X(x; \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

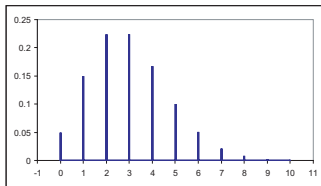
parameter $\lambda > 0$

kengetallen:

$$E(X) = \lambda$$

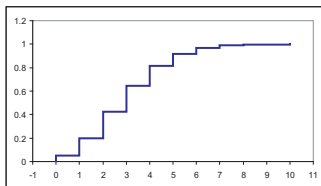
$$\text{var}(X) = \lambda$$

$$P(X = x)$$



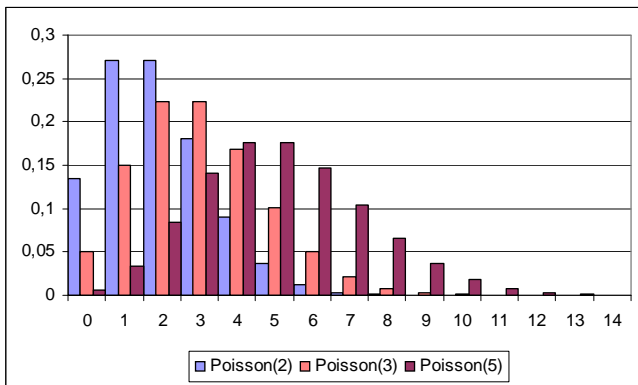
kansverdeling voor $\lambda = 3$

$$P(X \leq x)$$



cumulatieve verdelingsfunctie voor $\lambda = 3$

Voorbeeld: Poisson familie



Voorbeeld

- ▶ telefooncentrale
- ▶ 30 oproepen per uur
- ▶ kans op geen enkele oproep in 3 minuten tijd?
- ▶ eerst λ bepalen:

$$\begin{aligned} & 30 \text{ oproepen per uur} \\ & = 1.5 \text{ oproepen per 3 minuten} \end{aligned}$$

$$\Rightarrow \lambda = 1.5$$

- ▶ $P(X = 0) = p_X(0; 1.5) = ?$

Vervolg voorbeeld

- ▶ rekentoestel:

$$P(X=0) = \frac{(1.5)^0 e^{-1.5}}{0!} = 0.223$$

- ▶ R: `dpois(0, 1.5)`
Matlab: `poisspdf(0, 1.5)`
 - ▶ de waarde van x
 - ▶ de parameter λ

Vervolg voorbeeld

- ▶ kans op meer dan 5 oproepen in 5 minuten?
 $\Rightarrow \lambda = 2.5$
- ▶ $P(X > 5) = P(X \geq 6) = ?$
- ▶ $P(X \geq 6) = 1 - P(X \leq 5)$
R: `1 - ppois(5, 2.5)`
Matlab: `1 - poisscdf(5, 2.5)`
- ▶ rekentoestel:

$$\begin{aligned} P(X \geq 6) &= 1 - P(X = 0) - P(X = 1) - \dots - P(X = 5) \\ &= 1 - \frac{(2.5)^0 e^{-2.5}}{0!} - \dots - \frac{(2.5)^5 e^{-2.5}}{5!} \end{aligned}$$