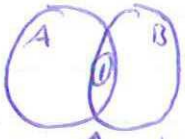
Statistiek

Hz. events + probability

- Intersection $A \cap B$, (i)

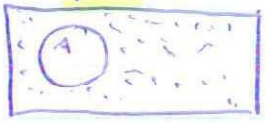


↑ → stukje overlap is intersection ($\cap$)

- union $A \cup B$

↳ zowel A als B, A+B, is de union, (U)

- Complement A^c



→ het complement is het tegenovergestelde van A, alles behalve A

- Demorgan's law

$$(A \cup B)^c = A^c \cap B^c \quad , \quad (A \cap B)^c = A^c \cup B^c$$

- Disjoint

→ A en B hebben geen overlap

- Probability function

$P(A)$ = kans op A

↳ Disjoint functies → $P(A \cup B) = P(A) + P(B)$

Vb. Hoe groot is de kans dat Piet gekozen wordt uit de groep: {Piet, Jan, Klaas, Pieter, Bart, Dirk}

- $\frac{1}{6}$ van de 6x wordt Piet niet gekozen

$P(\text{piet}) = \frac{1}{6}$

- Probability of an union

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- Probability of compliment

$$P(A^c) = 1 - P(A)$$

H3, Conditional Prob, multit bayes

- Conditional probability

Vb. - lange maanden: (jan, maa, mei, jul, aug, okt, dec) = 7
- maanden met r erin: (jan, feb, maar, apr, sep, okt, nov, dec) = 8

kans - $P(\text{lang}) = \frac{7}{12}$
- $P(r) = \frac{8}{12}$ → $P(R|L) = \frac{4}{7}$
↳ maanden met r die lang zijn

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

- multiplication rule

$$P(A \wedge B) = P(A|B) \cdot P(B)$$

- Bayes rule

$$P(C_i|A) = \frac{P(A|C_i) \cdot P(C_i)}{P(A|C_1)P(C_1) + P(A|C_2)P(C_2) + \dots}$$

Vb.
$$P(B|t) = \frac{P(T \wedge B)}{P(T)} = \frac{P(T|B) \cdot P(B)}{P(T|B)P(B) + P(T|B^c)P(B^c)}$$

Onafhankelijk?

$P(A|B) = P(A)$

$$P(A) = P(A|B) \cdot P(B) + P(A|B^c)P(B^c)$$

Ha,

- Bernoulli

Ber(p) → het gebeurt of het gebeurt niet

$P(X=1)$ = succes = p $\quad P(X=0)$ = failure = $1-p$

- Binomial

Bin(n,p) → het gebeurt of niet, In n-keer gevallen

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

p = kans

n = aantal maal

k = gevraagde

- geometric

Geo(p) $\quad P(x=k) = (1-p)^{k-1} \cdot p$

↳ $P(x>n) = 1-p^k$

Vb. bin

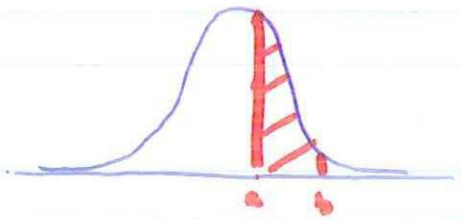
1000 lampen, kans dat lampje stuk is = 0,001 (P).

$P(X=0) = \binom{1000}{0} 0{,}001^0 (1-0{,}001)^{1000-0} = \ldots$

H5

- X. is continues

$$P(a \leq x < b) = \int_a^b f(x)\,dx$$



- uniform distributio.

Interval = $[\alpha, \beta]$

$$f(x) = \frac{1}{\beta - \alpha} \quad \text{for } \alpha \leq x \leq \beta$$

genoteerd: $U(\alpha, \beta)$

- exponential distributio.

$$f(x) = \lambda e^{-\lambda x} \qquad \lambda = \text{parameter}$$

$$F(a) = 1 - e^{-\lambda a}$$

genoteerd: $Exp(\lambda)$

- normal distributio.

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

genoteerd: $N(\mu, \sigma^2)$

H7, expectation, variance

- expectations

discrete

$E[X] = \sum a_i \cdot P(a_i)$

Vb.

a	2	4	6
$P(a)$	0,2	0,4	0,2

$E[X] = ((2 \cdot 0,2) + (4 \cdot 0,4) + (6 \cdot 0,2)) = 0,4 + 1,6 + 1,2 = 3,2$

continues

$E[X] = \int_{-\infty}^{\infty} x f(x) dx$

Vb. uniform distr. exs $[2,5] \rightarrow E[X] = \frac{\alpha+\beta}{2}$

$E[X] = \int_{-\infty}^{\infty} x \cdot f(x) dx = \int_{2}^{5} x \cdot \frac{1}{\beta - \alpha} dx = \int_{2}^{5} x \cdot \frac{1}{3} dx$

$\frac{1}{6}x^2 \Big]_2^5 = \frac{25}{6} - \frac{4}{6} = \frac{21}{6} = 3\frac{1}{2}$

Vb. geom.

$E[X] = \sum_{k=1}^{\infty} k \cdot p(1-p)^{k-1} = \frac{1}{p} \quad | \; Geo(p)$

Vb. exp

$E[X] = \frac{1}{\lambda} \quad | \; Exp(\lambda$

Vb. normal

$E[X] = \mu \quad | \; N(\sigma^2, \mu)$

- Variatie

$$\text{Var}[X] = E[X^2] - E[X]^2$$

Vb. normaal $\quad \text{Var}[X] = \sigma^2$

Vb. Exp $\quad \text{Var}[X] = 1/\lambda^2$

Vb. Uniform $\quad \text{Var}[X] = \frac{(\beta - \alpha)^2}{2}$

- Uitzonderingen:

$$E[rX + s] = r \cdot E[X] + s$$

$$\text{Var}[rX + s] = r^2 \cdot \text{Var}[X]$$

HS

- Change of units transformatie

$N(\mu, \sigma^2)$

Stel $P(X \leq 5)$ voor $N(4, 25)$

Dan moet deze kans eerst omgeschreven worden naar een $N(0,1)$!!

Daarvoor : $z = \frac{x - \mu}{\sigma}$

$z = \frac{x-4}{5}$ →

$P(x \leq 5) \rightarrow P(z \leq \frac{5-4}{5}) = P(z \leq 0,2) = 1 - 0,4207 = 0,5793$

- Jensen Inequality

$E[g(x)] = g(E[x])$

echter niet bij convex functies, 2^e afg. > 0

dan $g(E[x]) \leq E[g(x)]$

vb. $g(x) = x^2 \rightarrow g''(x) = 2$, convex

vb. $g(x) = e^{-x} \rightarrow g''(x) = e^{-x}$, convex

HG

- Joint probability

$P(a,b) = P(x=a, y=b)$

↳ Distribution

$F(a,b) = P(x \leq a, y \leq b)$

↳ continues distr.

$$P(a_1 \leq x \leq b_1 ; a_2 \leq y \leq b_2) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x,y)\, dx\, dy$$

Hio, covariance

- 2 dimensio. expectatio

discreet

$$E[g(x,y)] = \sum_i \sum_j g(a_i, b_j)\, p(x=a_i, y=b_j)$$

continues

$$E[g(x,y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)\, f(x,y)\, dx\, dy$$

vb.①	a: 0	1	2	$p(x=b)$
0	0	$\frac{1}{4}$	0	$\frac{1}{4}$
b: 1	$\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{1}{2}$
2	0	$\frac{1}{4}$	0	$\frac{1}{4}$
$p(x=a)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

$E[xy] = (0\cdot 0)\cdot 0 + (0\cdot 1)\cdot \frac{1}{4} + (2\cdot 0)\cdot 0 + \dots$ enzo

$= (a\cdot b)\cdot p(x=a, y=b)$

Let op! $E[x+y] = E[x] + E[y]$

- Covariance

$\cdot\ cov(x,y) = E[xy] - E[x]\cdot E[y]$

Vb①, gebruik tabel ①

$E[xy] = 1$, $E[x] = 1$, $E[y] = 1$

$cov = 1 - 1 = 0$

$\cdot\ var(x+y) = var(x) + var(y) + 2cov(x,y)$

$\cdot\ cov(rx+s, ty+u) = r\cdot t\cdot cov(x,y)$

- Density, coefficient

$$\rho(x,y) = \frac{cov(x,y)}{\sqrt{var(x)\cdot var(y)}}$$

Hfd 4 Poisson

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

$\lambda = \lambda \cdot t$ → aantal p. tijds eenheid

gerekend $Pois(\lambda)$

Vb. 12.6

er zit, in een koperen draad, om de 4 cm een fout

wat is de kans op 1 fout in een meter

λ = aantal verwachten fout in 1 meter = 2,5

$k = 1$

$$P(X=1) = \frac{2,5^1}{1!} \cdot e^{-2,5} = 0,21$$

$$P(X=2) = \frac{2,5^2}{2!} e^{-2,5} = 0,26$$

- Expectatie, variantie

$E[X] = \lambda$, $\text{var}[X] = \lambda$

H13, law of big numbers

- average

$$\bar{X}_n = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} = \text{average}$$

- expectation, variance

$$E[\bar{X}_n] = \mu \quad , \quad \text{Var}[\bar{X}_n] = \frac{\sigma^2}{n}$$

H14, central limit

$$Z_n = \sqrt{n} \cdot \frac{\bar{X}_n - \mu}{\sigma} = \frac{\bar{X}_n - E[\bar{X}_n]}{\sqrt{\text{Var}[\bar{X}_n]}}$$

Statistick

H15

- Histogram

bin width = $\frac{\text{number } x_j}{n}$ n = aantal elementen

bin height = $\frac{\text{numer } x_j}{n \cdot |B_1|}$ $|B_1|$ = bin width

H16

43, 43, 41, 41, 41, 42, 43, 58, 58, 41

- Sample mean

average: $\bar{x}_n = \frac{x_1 + x_2 + x_3 + \dots x_n}{n} = 45$

- Sample media

midde van alle → 42

- Sample variance

$$S_n^2 = \frac{1}{n} \cdot \{(x_i - \bar{x}_n)^2$$

↳ Sample Standard deviation

$$S_n = \sqrt{\frac{1}{n-1} \{(x_i - \bar{x}_n)^2}$$

- mad.

mad = median ($|x_1 - med|$, $|x_2 - med|$ ...)

stel 3 4 5 8 9 → med = 5

mad = med(2, 1, 0, 3, 4) = 2

Hig, estimators

- Schätzer

$T = e^{-\bar{x}_n}$

- schätzers: $\mu \to \bar{x}_n$

$\sigma^2 = S_n^2$ (zu H16)

H20, mse

- $E[\bar{x}_n] = \frac{E[x_1] + E[x_2] \dots + E[x_n]}{n}$

- first estimator, based on sample mean

$T_1 = 2\bar{x}_n + 1$

$E[T_1] = N$ = number

- second estimator, based on sample maximum

$T_2 = \frac{n+1}{n} \cdot \underbrace{m_n}_{\downarrow \max} - 1$

$E[T_2] = N$

Vb. $61 : 19 \cdot 56 \cdot 24 - 16$

$T_1 = 2 \cdot \bar{x}_n + 1 \to \bar{x}_n = 35{,}2 \to T_1 = 71{,}4$

$T_2 = 72{,}2 = \frac{5+1}{5} \cdot 61 - 1$

Hz1, likelihood.

Zie Smoker p. 313

$L(p) = C \cdot P(x_1=1)^{29} \cdot p(x_2=2)^{16} \cdot p(x_3=3)^{17} \cdot \ldots \cdot p(x_i>12)^{42}$

Geometric destribution

~~$L(p) = C \cdot p^{16} \cdot (1-p)^{16} \cdot p^{17} \cdot (1-p)^{17} \cdot \ldots \cdot p^{12}$~~

$L(p) = C \cdot p^{29} \cdot ((1-p)^2 p)^{16} \cdot ((1-p)p^2)^{17} \cdot \ldots ((1-p)^{12})^7$

$= C p^{93} \cdot (1-p)^{322}$

max likelihood

$l'(p) = C \cdot p^{92}(1-p)^{321}(93 - 415p) = 0$

$p=0,\ p=1,\ p=0{,}224$

$$L = c \cdot p(x=0,1)^{440} \cdot p(x=2)^{93} \cdot p(x=3)^{95} \cdot p(x=4)^{7} \cdot p(x=7)^{1}$$

$$= L \cdot \left(\frac{\mu^2}{2!} e^{-\mu}\right)^{440}$$

Hz2

kleinste - kwadraten - methode

Data: paren: $(x_1, y_1), (x_2, y_2), (x_n, y_n)$

statistisch: $Y_i = \alpha + \beta x_i + u_i$

$$\beta = \frac{n \cdot (\Sigma x_i y_i) - (\Sigma x_i) \cdot (\Sigma y_i)}{n \Sigma x_i^2 - (\Sigma x_i)^2}$$

$$\alpha = \bar{y}_n - \beta \cdot \bar{x}_n$$

Hz3

Betrouwbaarheids interval

$$N(0,1) \rightarrow z = \frac{\bar{x}_n - \mu}{\sigma_x / \sqrt{n}}$$

$P(-1,96 \leq z \leq 1,96) = 0,95 =$ 95% betrouwbaarheid

$$P\left(-1,96 \leq \frac{\bar{x}_n - \mu}{\sigma_x \cdot \sqrt{n}} \leq 1,96\right)$$

$$P\left(\bar{x}_n - 1,96 \cdot \frac{\sigma_x}{\sqrt{n}} \leq \mu \leq \bar{x}_n + 1,96 \frac{\sigma_x}{\sqrt{n}}\right)$$

Interval $\left(\bar{x}_n - 1,96 \frac{\sigma_x}{\sqrt{n}}, \bar{x}_n + 1,96 \frac{\sigma_x}{\sqrt{n}}\right)$ heet een 95% betrouwbaarheids interval.

σ niet bekend:

$$T_{(schatting)} = \frac{\bar{x}_n - \mu}{s_n/\sqrt{n}}$$ = Student gemiddel

H26.

hypothese:

$N(\mu, \mu)$

$\sigma = 2$

H_0 = nulhypothese
H_1 = alternatieve hypothese

Vb snelheidslimiet = 120 km/h
↳ toegelaten 5% is vaak bekeurd.
↳ 3 metingen

$H_0: \mu = 120$
$H_1: \mu \geq 120$

$$T = \frac{\bar{x}_n - 120}{2/\sqrt{3}} \rightarrow T \sim N(0,1)$$

$z_{0,05} = 1,645 = T$

$H_0: \mu = 120$
$H_1: \mu > 120$ als $t > 1,645$

Stel $\mu = 123$

$$P(T < 1,645 \mid \mu = 123) = \frac{120 - 123}{2/\sqrt{3}} + 1,645$$

$$= \Phi(0,95) =$$

Stel $\mu = 121$

$$P(T < 1,645 \mid \mu = 121) = \frac{120 - 121}{2/\sqrt{3}} + 1,645 = \Phi($$

unbiased estimator

↳ expectation μ : $\bar{x}_n = \frac{x_1 \cdots x_n}{n} \to E[x_i]$

↳ variance σ^2 : $S_n^2 = \frac{1}{n-1} \sum_i^n (x_i - x_n)^2 \to Var(x_i)$

Zuivere schatter

↳ gemiddelde waardes

$$\bar{x}_n = \frac{x_1 + \cdots x_n}{n}$$

(T_1)

$$E[\bar{x}_n] = \frac{1}{2}(N+1)$$

$$\hookrightarrow T = 2 \cdot \bar{x}_n - 1$$

$$\to Var[T] = Var[2 \cdot \bar{x}_n - 1] = 4 \cdot Var[\bar{x}_n] = \frac{(N+1)(N-n)}{3n}$$

↳ max genomen waarde

(T_2)

$$m_n = \max[x_1, x_2, \ldots x_n]$$

$$E[m_n] = \frac{n}{n+1}(N+1) \to Var[T] = \frac{(N+1)(N-n)}{n(n+2)}$$

$$T = \frac{n+1}{n} m_n - 1$$

$$\frac{Var(T_1)}{Var(T_2)} = \frac{n+2}{3}$$

T_2 = zuiverder wanneer er een getal > (groter) dan 1 uit komt

Mean squared error

$MSE = Var(T) - (E[T) - \theta)^2$

$= \text{Variatie} - \text{afwijk}^2$

Likelihood

bar $\rightarrow$ p of $1-p$

	1	2	3	4	5
kop	0	0	1	1	1
munt	3	3	2	2	2

~~$P(\bar{x}_n) = p^3 (1-p) \cdot (1-p) \cdot$~~

$= p^0 \cdot ((1-p)p]^0 \cdot ((1-p)^2 p)^1 \cdot ((1-p)^3 p)^1 \cdot ((1-p)^4 p)^1$

$= 1 \cdot 1 \cdot (1-p)^2 p \cdot (1-p)^3 p \cdot (1-p)^4 p$

$= (1-p)^2 p \cdot$

$= (1-p)^5 p^3$

MaD = mediaan van mediaan afw.

Stel: 42, 48, 50, 51, 57, 59, 62, 72

mediaan = 54

afw: 12, 6, 4, 3, 3, 4, 8, 18

: 3, 3, 4, 4, 6, 8, 12, 18

mediaan van mediaan afw

$MAD = \frac{1}{2}(4+6) = 5$

Steekproef afw.

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

$\bar{x}_n = 55$

$$s_8^2 = \frac{1}{7}\sum_{i=1}^{8}(x_i - 55)^2$$

$$= \frac{1}{7}\cdot((42-55)^2 + (48-55)^2 + \dots)$$

$$= \frac{1}{7}\cdot 610 = 87,142$$

$s = 9,34$

empirische verdelings functie

$F_5(60) = \frac{\text{aantal punten} \leq 60}{n} = \frac{6}{8} = 0,75$

Kans-massa-functie $S = a+b = x+y$

$P(a,b)$	b: 0	1	2	3
a: −1	0,05	0,15	0,05	0,1
0	0	0,05	0,15	0,05
2	0,05	0,2	0,05	0,1

→

$a+b$	a: −1	0	2
0	−1	0	1	2
1	0	1	2	3
2	2	3	4	5

s	−1	0	1	2	3	4	5
$P(s)$	0,05	0,15	0,1	0,3	0,25	0,05	0,1