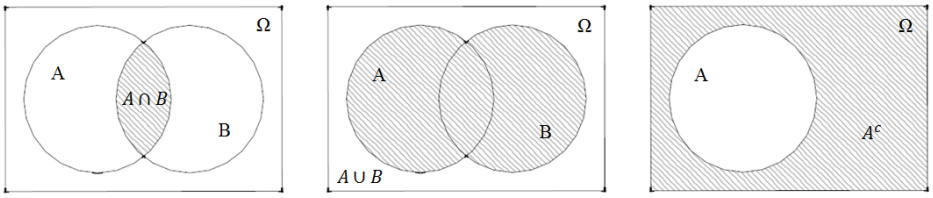
Kansrekening & Statistiek

Definitie
Een *kansruimte* is een verzameling Ω van *uitkomsten*, met een collectie deelverzamelingen $A_1, A_2, \ldots, B, C$: *gebeurtenissen* waarvoor *kansen* gedefinieerd zijn.

$$P(A) = "kans\ op\ A"$$



Definitie
Een *kansfunctie* voegt kansen toe aan gebeurtenissen, en heeft de volgende twee eigenschappen:

$$P: \{gebeurtenissen\} \to [0,1]$$

1. $P(\Omega) = 1$
2. Voor *disjuncte* gebeurtenissen A, B:
   $P(A \cup B) = P(A) + P(B)$
3. $P(\emptyset) = 0$
4. $P(A^c) = 1 - P(A)$
5. $P(A) = P(A \cap B) + P(A \cap B^c)$
6. $Voor\ disjuncte\ A_1, A_2, A_3, \ldots, A_k$:
   $P(A_1 \cup A_2 \cup \ldots \cup A_k) = P(A_1) + P(A_2) + \cdots + P(A_k)$
7. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ $\quad$ $(\boldsymbol{somregel})$

Wanneer de uitkomstensruimte Ω eindig veel uitkomsten $\omega_1, \ldots, \omega_n$ bevat met gelijke kansen:

$$P(\{\omega_i\}) = \frac{1}{n}, \qquad i = 1, \ldots, n$$

dan kun je de kans op een gebeurtenis A uitrekenen door te tellen:

$$P(A) = \frac{\#A}{\#\Omega} \qquad \left(= \frac{aantal\ elementen\ van\ A}{aantal\ elementenvan\ \Omega}\right)$$

Definitie
Voor twee gebeurtenissen A, B, met $P(B) \neq 0$ definiëren we de kans op A gegeven B door:

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

Hieruit volgt direct:

1. $P(A \cap B) = P(B)P(A|B)$ $\quad$ *de* $\boldsymbol{productregel}$ *voor kansen*
2. $P(A) = P(A \cap B) + P(A \cap B^c)$
   $= P(B)P(A|B) + P(B^c)P(A|B^c)$ $\quad$ *de* $\boldsymbol{wet\ van\ de\ totale\ kans}$ $(\boldsymbol{kansboom})$
3. $P(A \cap B \cap C) = P(A \cap B)P(C|A \cap B)$
   $= P(A)P(B|A)P(C|A \cap B),$ $\quad$ *enz.*
4. $Als\ C_1,\ C_2,\ C_3\ disjunct\ zijn\ en\ C_1 \cup C_2 \cup C_3 = \Omega, dan$:
   $P(A) = P(C_1)P(A|C_1) + P(C_2)P(A|C_2) + P(C_3)P(A|C_3),$ $\quad$ *enz.*

Stelling
Bayes' Rule

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)} = \frac{P(B)P(A|B)}{P(B)P(A|B) + P(B^c)P(A|B^c)}$$

Definitie
Twee gebeurtenissen A, B, met $P(B) \neq 0$ heten *onafhankelijk* als:

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

Stelling
Voor onafhankelijke gebeurtenissen A en B geldt:

$$P(A \cap B) = P(A)P(B)$$

Definitie
n gebeurtenissen $A_1, \dots, A_n$, heten onafhankelijk als voor elk k-tal geldt:

$$P(A_{i1} \cap A_{i2} \cap \dots \cap A_{ik}) = P(A_{i1})P(A_{i2}) \dots P(A_{ik})$$

Equivalent hiermee:

$$P(A_{i1}|A_{i2} \cap A_{i3} \cap \dots \cap A_{ik}) = P(A_{i1})$$

Onafhankelijkheid van een n-tal is niet het zelfde als paarsgewijze onafhankelijkheid.
In het algemeen is: $P(A \cap B) \neq P(A)P(B)$
Dit geldt namelijk alleen als A en B onafhankelijk zijn.

$P(A \cup B) = P(A) + P(B)$ $\quad geldt\ alleen\ als\ A\ en\ B\ disjunct\ zijn!$
$P(A \cap B) = P(A)P(B)$ $\quad geldt\ alleen\ als\ A\ en\ B\ onafhankelijk\ zijn!$

Definitie
Een *discrete stochast* is een functie $X: \Omega \to R$ op een kansruimte die hetzij eindig veel waarden $a_1, a_2, a_3, \dots$ aanneemt.
Intuïtief: een stochast X 'vertaalt' uitkomsten in getallen.

Definitie
De *kansmassafunctie* p is de functie $p: R \to [0,1]$ gedefinieerd via $p(a) = P(X = a)$.
Voor de meeste reële getallen a zal: $p(a) = 0$

Definitie
Een *verdelingsfunctie* van de (discrete) stochast X is de functie $F: R \to [0,1]$ gedefinieerd door $F(a) = P(X \leq a)$.
F is een niet-dalende functie met:

$$\lim_{x \to -\infty} F(x) = 0 \qquad en \qquad \lim_{x \to \infty} F(x) = 1$$

Definitie
Een *Bernoulli-variabele* is een stochast die alleen de waarden 0 en 1 aanneemt.
Een bernoulli-variabele 'vertaalt' de twee mogelijke uitkomsten van een 'experiment' – bijv. "falen"/"succes" – naar 0 / 1.
Definitie
Een *binomiale variabele met parameters n en p* is een stochast X die waarden $0,1,\dots,n$ aanneemt met kansen:

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

Een binomiale variabele 'hoort bij' het aantal successen in een serie van n (onafhankelijke) 'experimenten' met dezelfde 'succeskans' p.
Als $N \sim Bin(n,p)$ $\quad dan: N = 0,1,\dots,n$
N is alleen binomiaal als N elke waarde van $0, 1, \dots, n$ kan aannemen.

Definitie
Een *permutatie van (lengte) k* uit $\{a_1, a_2, \dots, a_n\}$ is een *geordend rijtje* $(a_{i1}, a_{i2}, \dots, a_{ik})$.
Het *aantal permutaties* van k uit $\{a_1, a_2, \dots, a_n\}$ is gelijk aan:

$$P(n,k) = n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot \big(n-(k-1)\big) = \frac{n!}{(n-k)!}$$

Een *combinatie van (lengte) k* uit $\{a_1, a_2, \dots, a_n\}$ is een *geordende set* $(a_{i1}, a_{i2}, \dots, a_{ik})$.
Het $aantal\ combinaties$ van k uit $\{a_1, a_2, \dots, a_n\}$ is gelijk aan:

$$C(n,k) = \frac{n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot (n-(k-1))}{k!} = \frac{n!}{k!\,(n-k)!} =_{def} \binom{n}{k}$$

Definitie
Een *geometrische variabele met parameter p* is een stochast X die de waarden 1,2,3,… aanneemt met kansen:

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

Een geometrische variabele 'hoort bij' het aantal (onafhankelijke) 'experimenten' met dezelfde 'succeskans' p, dat nodig is tot (en met) het eerste 'succes'.

Definitie
Een *continue stochast* is een stochast X waarvoor een functie f bestaat waarmee je kansen uit kunt rekenen via:

$$P(a \le X \le b) = \int_a^b f(x)dx$$

De functie f heet de *dichtheid* van X.
Een dichtheid is niet-negatief, en $\int_{-\infty}^{\infty} f(x)dx = 1$.
De verdelingsfunctie is exact zo gedefinieerd als bij discrete stochasten: $F(a) = P(X \le a)$.
Dat geeft het verband: $F(a) = \int_{-\infty}^{a} f(x)dx$
Omgekeerd geldt dan: $f(x) = F'(x)$.

Definitie
Als $0 \le p \le 1$ en X is een stochastische variabele, dan is het *p-de kwantiel* of *100p-de percentiel* het kleinste getal q_p zodat:

$$F(q_p) = P(X \le q_p) = p$$

Definitie
Een stochast X heeft een *uniforme verdeling op [a,b]* als:

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & als\ a \le x \le b \\ 0, & daarbuiten \end{cases} \quad en\ dus, \quad F(x) = \begin{cases} 0, & als\ x \le a \\ \dfrac{x-a}{b-a}, & als\ a \le x \le b \\ 1, & als\ x \ge b \end{cases}$$

Notatie: $U[a,b] \quad of \quad U(a,b)$
Een uniforme variabele beschrijft een 'willekeurig' getal in een interval $[a,b]$.

Definitie
De stochast X heeft een *normale verdeling met parameters μ en σ* ($\sigma > 0$), notatie $X \sim N(\mu, \sigma^2)$, als:

$$f(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}, \qquad x \in R \qquad Niet\ uit\ het\ hoofd\ leren!$$

Waarden van de verdelingsfunctie zijn getabelleerd (*normcdf*).

In de praktijk: som (of gemiddelde) van een groot aantal 'random' effecten – bijv. meetfouten.

Definitie
De stochast X heeft een *exponentiële verdeling met parameter* $\lambda(>0)$ als:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & als\ x \geq 0 \\ 0, & als\ x < 0 \end{cases} \quad en\ dus, \quad F(x) = \begin{cases} 0, & als\ x < 0 \\ 1 - e^{-\lambda x}, & als\ x \geq 0 \end{cases}$$

Beschrijft een 'geheugenloze' wachttijd: als $T \sim Exp(\lambda)$, dan:

$$P(T \geq s + t | T \geq t) = P(T \geq s)$$

Definitie
De *verwachting* van een discrete stochast X is gedefinieerd door:

$$E[X] = \sum_a a\, P(X = a) = \sum_a a\, p(a)$$

en voor een continue stochast met dichtheid f door:

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

De verwachting is te interpreteren als 'het gemiddelde op de lange duur'.

Stelling
Als $Y = g(x)$, dan geldt:

$$E[Y] = \sum_a g(a) P(X = a) \qquad of \qquad E[Y] = \int_{-\infty}^{\infty} g(x) f(x)\, dx$$

Gevolg: voor $g(x) = ax + b$, leidt dit tot: $E[aX + b] = aE[X] + b$

Definitie
De *variatie* van een stochast X is gedefinieerd door:

$$Var(X) = E[(X - E[X])^2]$$

De wortel hiervan heet de *standaardafwijking (of –deviatie).*

De variantie is een maat voor de 'spreiding' van een stochast X.

Stelling
1. $Var(X) = E[X^2] - (E[X])^2$
2. $Var(aX + b) = a^2\, Var(X)$

Discreet	$E[X]$	$Var(X)$
$Ber(p)$	p	$p(1-p)$
$Bin(n,p)$	np	$np(1-p)$
$Geom(p)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$

Continu	$E[X]$	$Var(X)$
$Unif(a,b)$	$\frac{1}{2}(a+b)$	$\frac{1}{12}(b-a)^2$
$N(\mu, \sigma^2)$	μ	σ^2
$Exp(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Transformeren van een discrete stochast
Stel de uitkomsten van X zijn $a_1, a_2, \ldots$ met kansen $p(a_1)$, wat is de verdeling van $Y = g(X)$?
De uitkomsten van Y: $g(a_1),\ g(a_2), \ldots$
Verder:

$$P(Y = b) = \sum_{\{a | g(a) = b\}} P(X = a) = \sum_{\{a | g(a) = b\}} p(a)$$

Transformeren van een continue stochast
Dit gaat het eenvoudigst via de verdelingsfunctie.
Stel stochast X heeft verdelingsfunctie F_X. Wat is de verdelingsfunctie van $Y = g(X)$?
Er geldt:

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = \cdots = P\big(X \leq g^{-1}(y)\big) = F_X(g^{-1}(y))$$

Stelling
Stel stochast X heeft een normale verdeling met parameters μ, σ^2, dan heeft $Z = \frac{X-\mu}{\sigma}$ een *standaard normale verdeling*.
Algemener:

$$Y = aX + b \quad heeft\ een \quad N(a\mu + b, a^2\,\sigma^2)$$

Als $\mu = E[X]$ en $\sigma^2 = Var(X)$ bestaan dat heet $Y = \frac{X-\mu}{\sigma}$ de gestandaardiseerde van X. Er geldt:

$$E[Y] = 0 \qquad en \qquad Var(Y) = 1$$

Een tabel voor de standaard normale verdeling volstaat. Bijvoorbeeld:

$$P(X \leq a) = P\left(\frac{X-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}\right) = P\left(Z \leq \frac{a-\mu}{\sigma}\right), \quad Z \sim N(0,1)$$

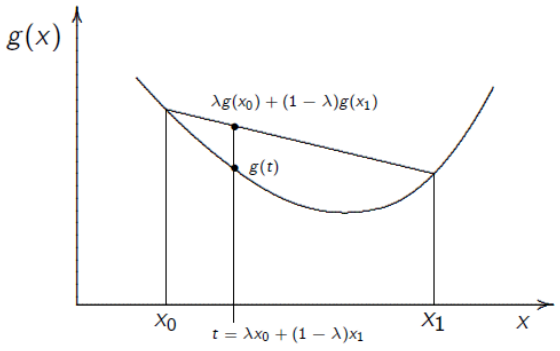
Maxima en minima van (onafhankelijke) stochasten
Stel $X_1, X_2, \ldots, X_n$ zijn onafhankelijke stochasten met dezelfde verdeling.
Laat $M = max\{X_1, X_2, \ldots, X_n\}$ en $Z = \min\{X_1, X_2, \ldots, X_n\}$

$$F_M(a) = F(a)^n \qquad en \qquad F_Z(a) = 1 - \big(1 - F(a)\big)^n$$

Definitie
De functie $f: [a,b] \to R$ is *convex*, als voor elk tweetal punten $x_0 < x_1$ in het interval (a,b), en elk getal $0 \leq \lambda \leq 1$ geldt:

$$f(\lambda x_0 + (1-\lambda)x_1) \leq \lambda f(x_0) + (1-\lambda) f(x_1)$$

Als de functie g twee keer differentieerbaar is, en $g''(x) \geq 0$ voor $x \in (a,b)$, dan is g convex op $[a,b]$.



Stelling (ongelijkheid van Jensen)
Stel stochast X neemt waarden aan in het interval $[a,b]$, en de functie g is convex op $[a,b]$. Dan geldt:

$$E[g(X)] \geq g(E[X])$$

Stelling
Stel F is een stijgende functie van het interval $[a,b]$ op het interval $[0,1]$, en U is uniform verdeeld op $[0,1]$. Dan heeft $X = F^{-1}(U)$ precies de verdelingsfunctie F.

$$P(X \leq a) = P(F^{-1}(U) \leq a) = P\big(U \leq F(a)\big) = F(a)$$

Simulatie wordt toegepast wanneer het precies berekenen te complex wordt of niet mogelijk is.

Definitie
De *gezamenlijke kansmassafunctie* van een paar stochasten X, Y is gedefinieerd door:

$$p(a,b) = P(X = a, Y = b), \qquad a, b \in R$$

Definitie
Een paar stochasten X, Y heeft een *gezamenlijke (kans)dichtheid* f als:

$$P\big((X,Y) \in G\big) = \int\int_G f(x,y)dx\,dy$$

Definitie
De *gezamenlijke verdelingsfunctie* is in beide gevallen gedefinieerd door:

$$F(a,b) = P(X \leq a, Y \leq b)$$

Stelling
Discreet: Hebben X en Y de gezamenlijke kansmassafunctie p, dan vind je de *(marginale) kansverdeling* van X hieruit terug via:

$$p_X(a) = P(X = a) = \sum_b P(X = a, Y = b) = \sum_b p(a,b)$$

en analoog door p_Y.

Continu: Hebben X en Y de gezamenlijke dichtheid $f(x,y)$, dan vind je de *(marginale) dichtheid* van X terug via:

$$f_X(a) = \int_{-\infty}^{\infty} f(a,y)dy$$

En analoog voor Y:

$$f_Y(b) = \int_{-\infty}^{\infty} f(x,b)dx$$

Stelling
Voor continue stochasten geldt de ene kant op:

$$F(x,y) = P(X \leq a, Y \leq y) = \int_0^x \int_0^y f(u,v)du\,dv$$

en de andere kant op:

$$f(x,y) = \frac{\partial^2 F}{\partial x \partial y}(x,y)$$

Definitie
$X_1, X_2, \dots, X_n$ heten *onafhankelijk* als voor alle $a_1, a_2, \dots, a_n$ geldt:

$$P(X_1 \leq a_1, \dots, X_n \leq a_n) = P(X_1 \leq a_1) \cdot \dots \cdot P(X_n \leq a_n)$$

Anders gezegd, als: $F(a_1, a_2, \dots, a_n) = F_{X_1}(a_1) \cdot F_{X_2}(a_2) \cdot \dots \cdot F_{X_n}(a_n)$

Equivalent hiermee: $X_1, X_2, \dots, X_n$ zijn onafhankelijk als voor de gezamenlijke dichtheid geldt:

$$f(a_1, a_2, \dots, a_n) = f_{X_1}(a_1) \cdot f_{X_2}(a_2) \cdot \dots \cdot f_{X_n}(a_n)$$

Stelling
Discreet: Als X de kansmassafunctie $p(a) = P(X = a)$ heeft, en $Y = g(X)$, dan:

$$E[Y] = \sum_i g(a_i)P(X = a) = \sum_i g(a_i)p(a_i)$$

Continu: Als X de dichtheid f heeft, en $Y = g(X)$, dan:

$$E[Y] = \int_R g(x)f(x)dx$$

Stelling
Discreet: Als X en Y de gezamenlijke kansmassafunctie $p(a,b)$ hebben, en $Z = g(X,Y)$, dan:

$$E[Z] = \sum_{i,j} g(a_i, b_j) P(X = a, Y = b) = \sum_{i,j} g(a_i, b_j) p(a_i, b_j)$$

Continu: Als X en Y de gezamenlijke dichtheid f hebben, en $Z = g(X,Y)$, dan:

$$E[Z] = \iint_{R^2} g(x,y) f(x,y) dx\, dy$$

Hierdoor geldt:

$$E[aX + bY + c] = aE[X] + bE[Y] + c$$

En geldt altijd:

$$Var(X+Y) = Var(X) + Var(Y) + 2E[(X - E[X])(Y - E[Y])]$$

Definitie
De *covariantie* van X en Y is gedefinieerd door:

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$
$$Cov(X,Y) = E[XY] - E[X]E[Y]$$

En dus:

$$Var(X+Y) = Var(X) + Var(Y) + 2\, Cov(X,Y)$$

X en Y heten *positief/negatief gecorreleerd* al naar gelang $Cov(X,Y) > 0 \;\; of \;\; < 0$.
Als $Cov(X,Y) = 0$, dan heten X en Y *ongecorreleerd*.

Stelling
Voor onafhankelijke stochasten X en Y geldt:

$$E[XY] = E[X]E[Y]$$

Gevolg: onafhankelijke stochasten hebben een covariantie van 0. Of te wel, onafhankelijke stochasten zijn altijd ongecorreleerd. Het omgekeerd geldt niet!

Stelling
- $Cov(aX + b, cY + d) = ac\, Cov(X,Y)$;
- $Cov(X, Y + Z) = Cov(X,Y) + Cov(X,Z)$.

Definitie
De *correlatie(coëfficiënt)* van X en Y is gedefinieerd door:

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

De correlatie zegt alleen iets over een mogelijk *lineair verband*.

Stelling
- $-1 \le \rho(X,Y) \le 1$
- $\rho(aX + b, cY + d) = \rho(X,Y)$, als $ac \ge 0$ en $-\rho(X,Y)$, als $ac \le 0$, m.a.w. $\rho(x,Y)$ is *schaalinvariant*.

Stelling
Als X en Y de gezamenlijke kansmassafunctie $p(a,b)$ hebben, dan heeft $S = X + Y$, de kansmassafunctie:

$$p_S(c) = \sum_{(i,j):a_i + b_j = c} p(a_i, b_j) = \sum_i p(a_i, c - a_i)$$

In het geval dat X en Y onafhankelijk zijn wordt dit:

$$p_S(c) = \sum_i p(a_i) p(c - a_i)$$

Stelling
De som van twee onafhankelijke binomiale variabelen met dezelfde parameter p, is weer een binomiale variabele.

$$\underbrace{Bin(n,p) + Bin(m,p)}_{Onafhankelijk} = Bin(n+m,p)$$

Stelling
Als X en Y de onafhankelijke stochasten zijn met dichtheden f_X en f_Y, dan wordt de dichtheid van $S = X + Y$ gegeven door:

$$f_S(c) = \int_{-\infty}^{\infty} f_X(t) f_Y(c-t)dt \qquad \left(= \int_{-\infty}^{\infty} f_X(c-t) f_Y(t)dt\right)$$

Stelling
De som S_n van n onafhankelijke exponentiële variabelen met dezelfde parameter λ heeft een $Gamma(n,\lambda)$-verdeling:

$$f_{S_n}(x) = \frac{\lambda(\lambda x)^{n-1} e^{-\lambda x}}{(n-1)!}$$

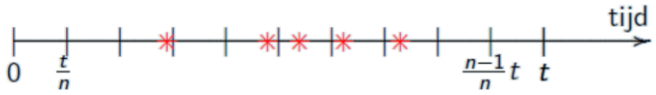
Stelling

- Als $X \sim N(\mu, \sigma^2)$, dan: $aX + b \sim N(a\mu + b, a^2\sigma^2)$
- Als $X \sim N(\mu_1, \sigma_1^2)$ en $Y \sim N(\mu_2, \sigma_2^2)$ onafhankelijke normale variabelen zijn, dan heeft $S = X + Y$ ook een normale verdeling.
  Uiteraard geldt dan: $S \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.
- Als Z_1 en Z_2 onafhankelijke standaard-normale variabelen zijn, en $U = a_1 Z_1 + a_2 Z_2$, $V = b_1 Z_1 + b_2 Z_2$, da hebben U en V een zogenoemde *bivariate normale verdeling*.
  Eenvoudig te bewijzen: $Cov(U,V) = a_1 b_1 + a_2 b_2$.

Definitie
Een rij gebeurtenissen op stochastische tijdstippen $X_1, X_2, X_3, \ldots$ is een *Poissonproces* als:

1. Het verwachte aantal gebeurtenissen in een interval van lengte u gelijk is aan λu; λ heet de *intensiteit*.
2. De aantallen gebeurtenissen N_1, N_2 in disjuncte tijdsintervallen zijn onafhankelijke stochasten.

Deel $[0,t]$ op in een groot aantal n deelintervallen.



Zij R_j het aantal aankomsten in het j-de deelinterval $I_{j,n}$ dan geldt bij benadering:

- $R_1, R_2, \ldots, R_n$ zijn onafhankelijk;
- R_j is 0 of 1. $P(R_j = 1) = \mathrm{E}[R_j] = \lambda \cdot$ lengte van $I_{j,n} = \lambda t/n$;
- $N_t = R_1 + \cdots + R_n$ heeft een $Bin(n,p)$ verdeling, $p = \lambda t/n$;

$$P(N_t = k) = \binom{n}{k}\left(\frac{\lambda t}{n}\right)^k \left(1 - \frac{\lambda t}{n}\right)^{n-k} \qquad voor\ k = 0, 1, \ldots, n$$

Het blijkt: een $Bin(n,p)$ verdeling, met n 'groot' en np 'klein' is bij benadering een Poisson-verdeling met parameter $\mu = np$.

Definitie
X heeft een *Poisson-verdeling met parameter* μ als:

$$P(X = k) = \frac{\mu^k}{k!} e^{-\mu} \qquad k = 1,2,3,\ldots$$
$$E[X] = \mu \quad en \quad Var(X) = \mu$$

Stelling
Stel $X_1, X_2, X_3, \ldots$ is een Poisson-proces met intensiteit λ, en laat $N(a,b)$ het aantal gebeurtenissen zijn in het interval (a,b). Dan:

1. $N(a,b)$ heeft een Poisson-verdeling met parameter $\mu = \lambda(b-a)$.
2. De tussentijden $T_1 = X_1, T_i = X_i - X_{i-1}, i = 2, 3, \ldots$ zijn onafhankelijk stochasten met een $Exp(\lambda)$ verdeling.
3. Gegeven dat er in een interval (a,b) precies k gebeurtenissen plaatsvinden, zijn de k tijdstippen waarop deze plaatsvinden onafhankelijk en uniform verdeeld in het interval (a,b).

Stelling
Stel $X_1, X_2, \ldots, X_n$ zijn onafhankelijke stochasten met dezelfde verwachting μ en dezelfde variantie σ^2, en laat $\bar{X}_n$ het gemiddelde zijn:

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

Dan gelden:

- $E[\bar{X}_n] = \mu$
- $Var(\bar{X}_n) = \frac{\sigma^2}{n}$

Stelling
Voor elke stochast X met een verwachting μ en variantie σ^2 geldt:

$$P(|X - \mu| > a) \leq \frac{\sigma^2}{a^2} = \frac{Var(X)}{a^2}$$

Dit is de *Chebyshev's ongelijkheid*.
Gevolg:

$$P(|X - \mu| > k\sigma) \leq \frac{\sigma^2}{(k\sigma)^2} = \frac{1}{k^2}$$

Stelling
Stel $X_1, X_2, \ldots, X_n$ zijn onafhankelijke stochasten met dezelfde verwachting μ en dezelfde variantie σ^2, en laat $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ en $\varepsilon > 0$. Dan geldt:

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0$$

$$E[\bar{X}_n] = \mu \qquad en \qquad Var(\bar{X}_n) = \frac{\sigma^2}{n}, \quad dus:$$

$$P(|\bar{X}_n - \mu| > \varepsilon) \leq \frac{Var(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0 \; als \; n \to \infty$$

Stelling
Stel $X_1, X_2, \ldots, X_n$ is een rij onafhankelijke stochasten met dezelfde verdeling.
Laat $F_n = \frac{aantal\ van\ de\ eerste\ n\ waarvoor\ a \leq X_i \leq b}{n}$, de *relatieve frequentie* zijn de van gebeurtenis "$X_i\ ligt\ in\ [a,b]$", en laat $p = P(a \leq X \leq b)$.
Dan geldt, voor $\varepsilon > 0$:

$$\lim_{n\to\infty} P(|F_n - p| > \varepsilon) = 0$$

M.a.w. F_n ligt voor 'grote' n 'dichtbij' $P(a \leq X \leq b)$.
Je kunt F_n schrijven als: $\frac{R_1 + R_2 + \cdots + R_n}{n}$, waarbij de R_i onafhankelijke Bernoulli stochasten zijn.

Stelling
Stel $X_1, X_2, X_3, \dots, X_n$ zijn onafhankelijke stochasten met dezelfde verdeling, met verwachting μ en variantie σ^2, en laat $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$. Dan geldt:

$$\lim_{n\to\infty} P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq a\right) = P(Z \leq a)$$

waarbij Z een standaard normale verdeling heeft.

$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ is de gestandaardiseerde van $\bar{X}_n$, dus heeft sowieso verwachting 0 en variantie 1.

Aangezien $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}S_n$, geldt:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\frac{S_n}{n} - \mu}{\sigma/\sqrt{n}} = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{S_n - E[S_n]}{\sqrt{Var(S_n)}}$$

Gestandaardiseerde van het gemiddelde = Gestandaardiseerde van de som.
Dus geldt ook:

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq a\right) = P(Z \leq a)$$

Stelling
Stel X heeft een $Bin(n, p)$ verdeling, met n 'groot' en p 'niet heel klein'. Dan geldt:

$$P(X \leq k) = P\left(\frac{X - np}{\sqrt{np(1-p)}} \leq \frac{k - np}{\sqrt{np(1-p)}}\right)$$
$$\approx P\left(Z \leq \frac{k - np}{\sqrt{np(1-p)}}\right)$$

waarbij Z weer een standaard normale verdeling heeft.

Voor de discrete variabele X geldt $P(X \leq k) = P(X < k + 1)$, maar:

$$P\left(Z \leq \frac{k - np}{\sqrt{np(1-p)}}\right) \neq P\left(Z \leq \frac{k + 1 - np}{\sqrt{np(1-p)}}\right)$$

De 'gulden middenweg' geeft een betere benadering:

$$P(X \leq k) \approx P\left(Z \leq \frac{k + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$

Dit wordt wel de *continuïteitscorrectie* genoemd.

Methoden om een *dataset* van metingen samen te vatten

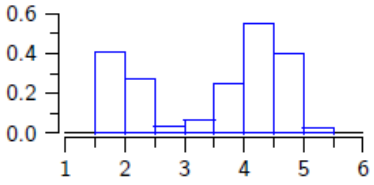
- Grafisch
    1. Histogram;
    2. Kern(dichtheids)schatting;
    3. Empirische verdelingsfunctie;
    4. Scatterplot.
- Numeriek
    1. Locatie
    2. spreiding;
    3. Empirische kwantielen;
    4. Boxplot.

Histrogram

Verdeel het gebied waarover $x_1, \dots, x_n$ verspreid liggen in *cellen* $B_1, \dots, B_m$.

Op cel B_i is de hoogte:

$$h_i = \frac{factie\ van\ de\ x_j's\ die\ in\ B_i\ zitten}{breedte\ van\ B_i} = \frac{het\ aantal\ x_j's\ in\ B_i}{n|B_i|}$$



Idee: *oppervlakte* van de rechthoekje op cel B_i is:

$$\frac{het\ aantal\ x_j's\ in\ B_i}{n}$$

Is de *relatieve frequentie* van de gebeurtenis 'tijdsduur lig in B_i.

Is een schatting van de *'kans'* om in cel B_i terecht te komen.

Kern(dichtheid)schatting

Idee: op elk datapunt x_i een 'kanshoopje' $1/n$ in de vorm van $K(x)$.

Kernfunctie $K(x)$:

- Een niet-negatieve functie;
- $\int_{-\infty}^{\infty} K(x)dx=1$;
- Symmetrisch t.o.v. 0, d.w.z. $K(-x)=K(x)$;
- Meestal: $K(x)=0$ buiten het interval [-1,1].

Kenschatting:

$$f_{n.h}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-x_i}{h}\right)$$

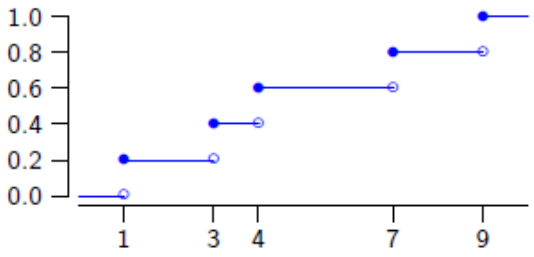
Empirische verdelingsfunctie

$$F_n(a) = \frac{aantal\ elementenin\ de\ dataset \leq a}{n}$$

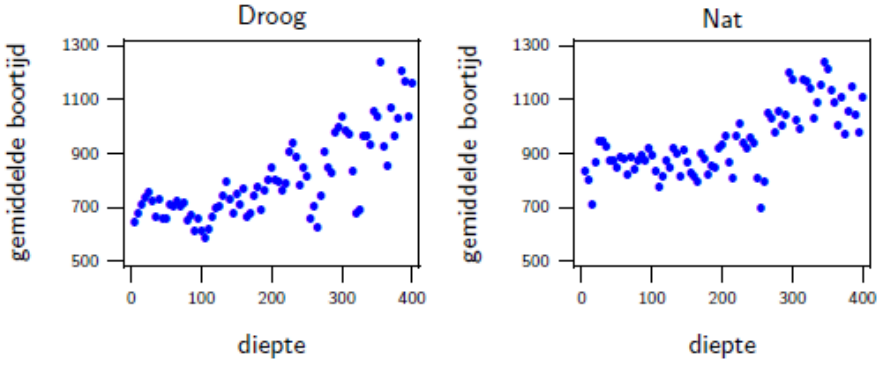
Kortom: Schatting voor 'de kans om op of onder a terecht te komen'.

Empirische verdelingsfunctie voor dataset 1 3 4 7 9:

Hoe steiler hoe dichter de elementen van de dataset bij elkaar ligger.



Scatterplot



Aanduiding van de 'locatie' van een dataset
(Steekproef)gemiddelde:

$$\bar{x}_n = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

(Steekproef)mediaan:
$Med(x_1, x_2, \ldots, x_n)$ = 'middelste' element naar volgorde van grootte.
(Bij even aantal punten: gemiddelde van de twee 'middelste'.)
Het gemiddelde is gevoelig voor uitschieters.

Maten voor de spreiding
Steekproefvariantie

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

Steekproefstandaarddeviatie

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

Mediane absolute deviatie (MAD)

$$MAD(x_1, x_2, \ldots, x_n) = Med(|x_1 - Med_n|, \ldots, |x_n - Med_n|)$$

Mean of absolute deviations

$$mad(x_1, x_2, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}_n|$$

Empirische kwantielen
Dataset geordend naar volgorde van grootte:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

We stellen $x_{(i)}$ gelijk aan het $\frac{i}{n+1}$-de empirische kwantiel.
Voor een willekeurige $0 < p < 1$ is het *p-de empirische kwantiel*:

$$q_n(p) = x_{(k)} + \alpha(x_{(k+1)} - x_{(k)})$$

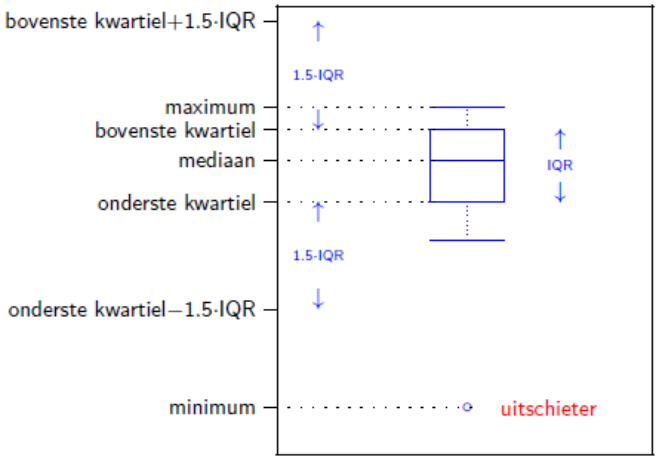
waarbij we het gehele getal k zo moeten kiezen dat geldt:

$$p = \frac{k}{n+1} + \frac{\alpha}{n+1}, \quad met\ 0 \leq \alpha \leq 1$$

Lineaire interpolatie
Dat wil zeggen: interpoleren tussen het k-de en $(k+1)$-ste element in de geordende dataset.
Waarbij: $k = \lfloor p(n+1) \rfloor$ en $\alpha = p(n+1) - k$
Hierbij:
$\lfloor x \rfloor$ = "x *naar beneden afgerond*".
Het 25% kwantiel $q_n(0.25)$ heet onderste (of: eerste) kwartiel.
Het 75% kwantiel $q_n(0.75)$ heet bovenste (of: derde) kwartiel.

De onderlinge afstand heet (inter)kwartielafstand:

$$IQR = q_n(0.75) - q_n(0.25)$$



Definitie
Een *steekproef* is een rij onafhankelijke stochasten $X_1, X_2, \dots, X_n$ met dezelfde verdeling.

Een steekproefgrootheid (ook wel: statistiek) is een stochast die een functie is van $X_1, X_2, \dots, X_n$, m.a.w. $Y = h(X_1, \dots, X_n)$.

Steekproefgrootheden kun je gebruiken om modelparameters te schatten:

(steekproef)gemiddelde $\bar{X}_n$	$\leftrightarrow$	verwachting
Steekproefmediaan	$\leftrightarrow$	mediaan
Empirische vdf $F_n(a)$	$\leftrightarrow$	verdelingsfunctie $F(a)$
Histogram	$\leftrightarrow$	dichtheid
Enz.		

Schatters
Setting
Data (metingen, observaties, ...) $x_1, x_2, \dots, x_n$ worden geacht afkomstig te zijn van een context die wordt beschreven door een *model* met een of meer onbekende *parameters*.

Definitie
Een *schatter* is een stochast/steekproefgrootheid $T = h(X_1, \dots, X_n)$ om een parameter θ te schatten.

Definitie
Een schatter T heet *zuiver (unbiased)* voor een parameter θ als $E[T] = \theta$.

Stelling
- Het steekprefgemiddelde, $T = \bar{X}_n$, is altijd een zuivere schatter voor de verwachting $\mu = E[X]$.
- De steekproefvariantie, $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$, is altijd een zuivere schatter voor de variantie $\sigma^2 = Var(X)$.

De steekproefstandaardafwijking is niet een zuivere schatter voor de standaarddeviatie van de verdeling van X.

De kwaliteit van een schatter
Intuïtief: in het geval van twee zuivere schatters T_1 en T_2:
T_2 is 'beter' dat T_1 als de variantie van T_2 kleiner is dan de variantie van T_1.

Definitie
Stel T_1 en T_2 zijn twee zuivere schatters voor een parameters θ. De *relatieve efficiëntie* van T_2 t.o.v. T_1 is gedefinieerd door:

$$\frac{Var(T_2)}{Var(T_1)} \qquad als\ Var(T_2) < Var(T_1)\ dan\ heet\ T_2\ efficiënter.$$

Definitie
De *mean squared error* van een schatter T (voor θ) is gedefinieerd door:

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

Stelling

$$MSE(T) = Var(T) + (E[T-\theta])^2 = Var(T) + (bias)^2$$
$$Bias\ =\ E[T] - \theta$$

Maximum likelihood principe
Kies de parameter(s) van een (vooraf bepaald) model, waarvoor de daadwerkelijke verkregen data de grootste kans van optreden heeft.

Dit principe geeft een zeer algemeen toepasbare methode, die bovendien schatters oplevert die goede of zelfs optimale eigenschappen hebben.

Stel: een dataset kan gezien worden als een realisatie van een steekproef $X_1, \dots, X_n$ uit een discrete verdeling, met een kansmassafunctie $p_\theta(x) = P_\theta(X = x)$, θ een onbekende parameter.
We definiëren de *likelihood* $L(\theta)$ van een dataset $x_1, \dots, x_n$ als:

$$L(\theta) = P(X_1 = x_1, \dots, X_n = x_n) = p_\theta(x_1) \cdot \dots \cdot p_\theta(x_n)$$

De *maximum likelihood schatting* van θ is die waarde die de likelihood $L(\theta)$ maximaliseert.
De bijbehorende schatter heet de *maximum likelihood schatter*.

Definitie
We definiëren in het continue geval de *likelihood* van een dataset $x_1, \dots, x_n$ als:

$$L(\theta) = f_\theta(x_1) f_\theta(x_2) \cdot \dots \cdot f_\theta(x_n)$$

En weer is de *maximum likelihood schatting* van θ die waarde die $L(\theta)$ maximaliseert.

De likelihoodfunctie $L(\theta)$ neemt voor de zelfde $\hat{\theta}$ een maximum aan als de *log-likelihoodfunctie* $l(\theta) = \ln(L(\theta))$.

Voor data $x_1, \dots, x_n$ uit $Unif(0, \theta)$-verdeling (met onbekende θ):

$$L(\theta) = \begin{cases} \frac{1}{\theta^n}, \theta \geq \max(x_1, \dots, x_n) \\ 0, \quad \theta < \max(x_1, \dots, x_n) \end{cases}$$

De maximum likelihood schatting voor θ is dan:

$$max(x_1, \dots, x_n)$$

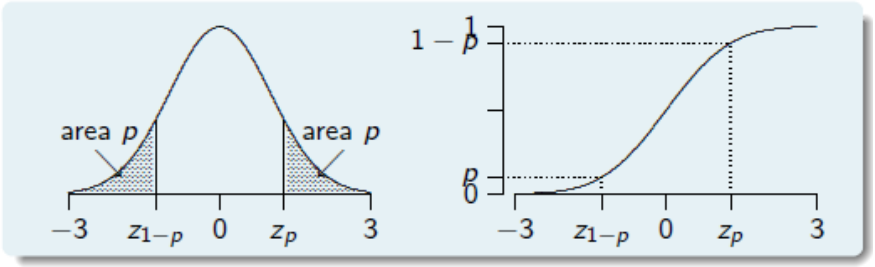
Stel: T is de maximum likelihood (ML) schatter voor θ, en $g: R \to R$ een of andere functie. Dan geldt:
$g(T)$ is de ML schatter voor $g(\theta)$.

Als T een zuivere schatter is voor θ, dan is $g(T)$ i.h.a. **niet** een zuivere schatter voor $g(\theta)$.

De maximum likelihood schatter is voor grote n praktisch zuiver en heeft de kleinst mogelijke variantie.

Betrouwbaarheidsintervallen
Setting
Op grond van data $x_1, \dots, x_n$ wordt een schatting gegeven van een parameter θ.
Gevraagd: een 'betrouwbaarheid' van de schatting.

Definitie
De kritieke waarde z_α is het getal waarvoor $P(Z \geq z_\alpha) = \alpha$, waarbij $Z \sim N(0,1)$. In voorbeeld $z_{0.025} = 1.96$.



Definitie
Dataset $x_1, \dots, x_n$ gegeven, gemodelleerd als realisatie van een steekproef $X_1, \dots, X_n$.
θ: de onbekende parameter.
$\gamma = 1 - \alpha$: een getal tussen 0 en 1, bijvoorbeeld 0.95.
Als er steekproefgrootheden $L_n = g(X_1, \dots, X_n)$ en $U_n = h(X_1, \dots, X_n)$ bestaan met:

$$P(L_n < \theta < U_n) = \gamma$$

voor elke waarde θ, dan heet (l_n, u_n) met $l_n = g(x_1, \dots, x_n)$ en $u_n = h(x_1, \dots, x_n)$ een $100\gamma\%$ betrouwbaarheidsinterval voor θ.
Het getal γ heet de betrouwbaarheid (of: betrouwbaarheidsniveau).

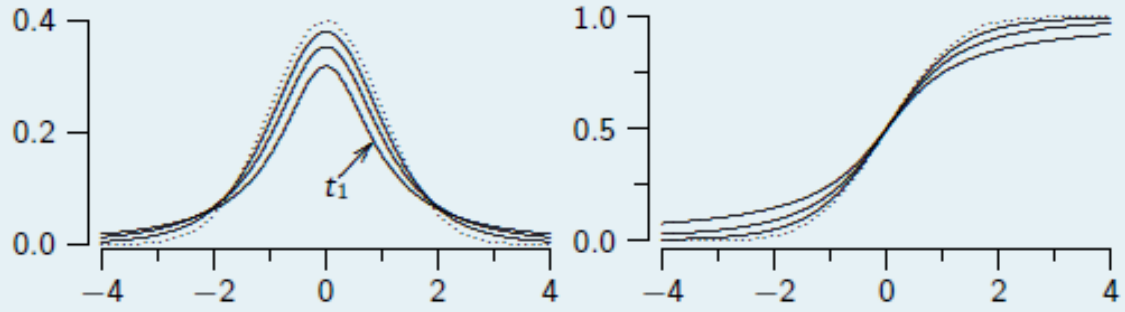
Interpretatie betrouwbaarheidsinterval: de **methode** waarmee het interval (l, u) is berekend, geeft in $xx\%$ van de gevallen een interval dat de parameter θ bevat.

Betrouwbaarheidsintervallen voor de verwachting μ

1. *Normale data, bekende σ:*
   Een $100(1-\alpha)\%$ betrouwbaarheidsinterval voor μ wordt gegeven door:
   $$\left(\bar{x}_n - z_{\frac{1}{2}\alpha}\frac{\sigma}{\sqrt{n}},\ \ \bar{x}_n + z_{\frac{1}{2}\alpha}\frac{\sigma}{\sqrt{n}}\right)$$
   Als de data niet perse normaal verdeeld zijn, maar n is groot, is de methode bij benadering OK, vanwege de centrale limietstelling.
2. *Normale data, onbekende σ:*
   Een $100(1-\alpha)\%$ betrouwbaarheidsinterval voor μ wordt gegeven door:
   $$\left(\bar{x}_n - t_{n-1,\frac{1}{2}\alpha}\frac{s_n}{\sqrt{n}},\ \ \bar{x}_n + t_{n-1,\frac{1}{2}\alpha}\frac{s_n}{\sqrt{n}}\right)$$
   waarbij $t_{n-1,\alpha}$ de *kritieke waarde* is van de *Student-verdeling* met $n-1$ vrijheidsgraden.
   Hier is het essentieel dat de data normaal verdeeld zijn.

Stelling
Als $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, onafhankelijk, dan heeft $\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$ een Student-verdeling met $n-1$ vrijheidsgraden.
Naarmate N groter wordt zal deze meer en meer op de $N(0,1)$-verdeling gaan lijken.
Voor betrouwbaarheidsintervallen heb je alleen de kritieke waarden nodig.

$$f(x) = k_m\left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}}, \quad k_m = \Gamma\left(\tfrac{m+1}{2}\right) / \left(\Gamma\left(\tfrac{m}{2}\right)\sqrt{m\pi}\right)$$



De $t(1)$, $t(2)$ en $t(5)$ verdeling, en de standaard normale (stippellijn).
Voor kleine m heeft $t(m)$ de dikste staarten, voor grote m is hij vrijwel gelijk aan de standaard normale verdeling.

Betrouwbaarheidsintervallen voor populatiefractie p
Stel een fractie p in een populatie heeft eigenschap A.
Data: een steekproef van grootte n levert k keer eigenschap A.
Schatting voor p: $\hat{p} = \frac{k}{n}$
Gevraagd: een $100\gamma\%$ betrouwbaarheidsinterval voor p.
Stel X is het aantal elementen uit een steekproef (van grootte n) met eigenschap A.
Dan: $X \sim Bin(n, p)$. De centrale limietstelling geeft:

$$\frac{X - np}{\sqrt{np(1-p)}} = \frac{X/n - p}{\sqrt{p(1-p)/n}} \approx N(0,1)$$

Dus:

$$P\left(-z_{\frac{1}{2}\alpha} \leq \frac{X/n - p}{\sqrt{p(1-p)/n}} \leq z_{\frac{1}{2}\alpha}\right) \approx 1 - \alpha$$

Eenzijdige betrouwbaarheidsintervallen
Gevraagd: op grond van data $x_1, \ldots, x_n$ uit een normale verdeling wordt een betrouwbaarheidsinterval (l, ∞) gevraagd voor de verwachting μ.
Gewenst: $L = g(X_1, \ldots, X_n)$ zodat $P(L \leq \mu) = \gamma = 1 - \alpha$.
Simpel idee: construeer een 'symmetrisch' $100(1-2\alpha)\%$ betrouwbaarheidsinterval, en 'laat de bovengrens weg'.
Dit 'werkt' vanwege de symmetrie van den normale (en ook: de Student-) verdeling:

$$P(L \leq \mu \leq U) = 1 - 2\alpha, \quad P(\mu \geq U) = \alpha \quad \rightarrow \quad P(L \leq \mu) = 1 - \alpha$$

Het toetsen van hypothesen
Nulhypothese en Alternatieve hypothese:
Beweringen over een model (van 'de werkelijkheid').

Type 1 fout:
De Nulhypothese wordt **verworpen** (ten gunste van de alternatieve hypothese), terwijl deze in feite **waar** is.

Type 2 fout:
De nulhypothese wordt **niet verworpen**, terwijl deze in feite **onwaar** is.

	Nulhypothese is waar	Nulhypothese is onwaar
Nulhypothese niet verworpen	OK	Fout van de tweede soort
Nulhypothese wordt verworpen	Fout van de eerste soort	OK

Toetsingsgrootheid:
Een steekproefgrootheid op grond waarvan besloten wordt wel of niet te verwerpen.

De *p-waarde* van een realisatie t_0 van de toetsingsgrootheid T:
De kans op een **minstens zo afwijkende** waarde van T, **gegeven dat de nulhypothese waar is**.
Afhankelijk van de alternatieve hypothese zal dit in het algemeen $P(T \geq t_0)$ of $P(T \leq t_0)$ zijn, maar 'tweezijdig' kan ook.

Definitie
Vaak wordt **van tevoren** afgesproken bij welke p-waarde de nulhypothese wordt verworpen. De maximale p-waarde waarbij verworpen wordt heet het *significantieniveau* α.
Dus: bij een p-waarde kleiner dan α: verwerp H_0.
bij een p-waarde groter dan α: verwerp H_0 niet.

Het *significantieniveau* is de maximale toegestane fout van type 1.

Het *kritieke gebied* van een toets (van H_0 versus H_1) is de verzameling uitkomsten K waarbij H_0 verworpen zal worden:
H_0 wordt verworden als de geobserveerde waarde t in K ligt.
De grenzen van het kritieke gebied heten de *kritieke waarden*.

Stappen voor het toetsen van hypothese:

1. Er is een **Onderzoeksvraag**
2. Er worden **Data** verzameld
3. Op grond van de data en de context een **Model** formuleren
4. De onderzoeksvraag vertalen naar **Hypothesen** over (parameter(s) van) **het model**
5. Kies een geschikte **Toetsingsgrootheid T**
6. Bepaal de **p-waarde** bij de geobserveerde waarde t van T, of:
   bereken het **Kritieke Gebied K**, en ga na of $t \in K$
7. **Verwerp al of niet** de nulhypothese
8. Geef de **Conclusie** met betrekking tot de onderzoeksvraag