TECHNISCHE HOGESCHOOL EINDHOVEN
Afdeling Algemene Wetenschappen
Onderafdeling der Wiskunde

WISKUNDE IV B

Waarschijnlijkheidsrekening en Statistiek

Syllabus van het College van

Prof. Dr. J.F. Benders

Gegeven in het Voorjaarssemester 1966

Bibl Mag

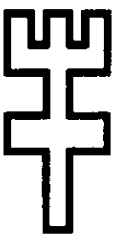
Onderafdeling der Wiskunde

Wiskunde IV B

Waarschijnlijkheids rekening en Statistiek

SYLLABUS VAN HET COLLEGE VAN
PROF. DR. J.F. BENDERS

GEGEVEN IN HET
VOORJAARSSEMESTER 1966



TECHNISCHE HOGESCHOOL EINDHOVEN

Onderafdeling der Wiskunde

W I S K U N D E I V B

Waarschijnlijkheidsrekening en Statistiek

Syllabus van het college

van

Prof. dr. J. F. Benders

gegeven in het

voorjaarssemester 1966

T e c h n i s c h e H o g e s c h o o l E i n d h o v e n

INHOUD

WAARSCHIJNLIJKHEIDSREKENING EN STATISTIEK

1. Inleiding

Een variabele is een grootheid welke een waarde kan aannemen uit een gegeven uitkomstengebied.

Zo'n uitkomstengebied kan van velerlei aard zijn, bv.: een verzameling gehele getallen, een stelsel vectoren, een gebied van punten in een n-dimensionale ruimte, een verzameling kleuren, enz..

Een variabele + uitkomstenruimte wordt echter pas interessant als vaststaat hoe, in een concrete situatie, aan die variabele een waarde uit het uitkomstengebied wordt toegekend.

Vele formules voorkomend in wiskundige modellen van fysische verschijnselen zijn typische voorbeelden van voorschriften hoe aan een variabele een waarde moet worden toegekend. Wil men bv. de hoogte h aangeven van een kogel boven een horizontaal vlak t seconden nadat deze met beginsnelheid V onder een hoek α is weggeschoten, dan luidt het voorschrift:

als $0 \leq t \leq \frac{2V}{g} \sin \alpha$ dan is $h = Vt \sin \alpha - \frac{1}{2}gt^2$;

als $t \geq \frac{2V}{g} \sin \alpha$ dan is $h = 0$.

Een belangrijke eigenschap van dit voorschrift is dat bij herhaald schieten "onder identieke omstandigheden", dus met dezelfde beginsnelheid V_0 en hoek α, en bij gelijk tijdsverloop t sedert het afschieten van de kogel, aan de variabele h dezelfde waarde wordt toegekend.

Een geheel andere situatie treffen we aan bij het dobbelspel. We spreken af dat speler A aan speler B een bedrag 100 moet betalen als bij een worp met de dobbelspelen een even getal boven komt en dat A van B een bedrag 50 ontvangt als een oneven getal boven komt.

Als variabele nemen we de opbrengst x van het spel voor speler A.

De uitkomstenruimte van x bestaat uit de twee getallen -100 en + 50.

Het toekennen van een waarde aan x gebeurt volgens het voorschrift:
doe een worp met de dobbelsteen;
geef x de waarde - 100 als een even getal boven komt;
geef x de waarde + 50 als een oneven getal boven komt.

Herhalen we dit spel "onder identieke omstandigheden" dan is het helemaal niet zeker dat aan x dezelfde waarde wordt toegekend. De waarde die x krijgt hangt af van het toeval; een kanselement, ook stochastisch element genoemd, speelt nu een rol. De variabele x wordt daarom ook een "stochastische variabele" genoemd.

Een volledig bevredigende definitie van "stochastische variabele" is slechts in abstract wiskundig verband te geven. Voor ons inleidend doel is de volgende werkdefinitie echter voldoende.

Definitie: een variabele heet stochastisch als de wijze waarop aan deze variabele in een concrete situatie een waarde uit de bijbehorende uitkomstenruimte wordt toegekend, een kanselement bevat.

Naar nederlands gebruik zullen we een stochastische variabele steeds met een onderstreepte letter aanduiden: $\underline{x}$, $\underline{a}$, $\underline{A}$ enz.. Men zij er echter op bedacht dat deze notatie niet algemeen gebruikelijk is en dat men in de literatuur vaak alleen uit het zinsverband kan opmaken of een variabele al dan niet stochastisch is.
Naast de aanduiding: stochastische variabele zullen we ook, zonder betekenisverschil, de in de literatuur voorkomende benamingen kansvariabele, toevalsvariabele, stochastiek, random variabele en aselecte grootheid, aantreffen.
Stochastische variabelen treft men in vele situaties aan. Soms is, evenals in bovenstaand voorbeeld, het kanselement expliciet bekend (dobbelspel, loterij, kaartspelen, roulette). Soms weten we ook alleen maar dat kanselementen een rol spelen en staan we juist voor het probleem over de aard daarvan wat meer informatie te krijgen. Dit laatste is het geval bij de meeste metingen: een meetresultaat wordt meestal beinvloed door toevallige storingen (niet precies instelbare temperatuur, aflees-

onnauwkeurigheid van meettoestellen, onzuiverheid van gereedschap enz.). Een voorschrift hoe een meting of waarneming verricht moet worden is daarom te beschouwen als een voorschrift hoe aan een bij de situatie passende stochastische variabele een waarde uit zijn uitkomstengebied moet worden toegekend.

Een meting of waarneming heeft tot doel te komen tot bepaalde conclusies, beslissingen en acties. Het stochastische karakter van een meting bemoeilijkt echter het trekken van conclusies uit het meetresultaat. Het is nu de wiskundige statistiek die, uitgaande van een wiskundige precisering van het intuitieve kansbegrip, aangeeft hoe waarnemingsgegevens geanalyseerd moeten worden om tot verantwoorde conclusies te komen.

Het is de bedoeling van dit college iets nader in te gaan op de wiskundige formulering van het kansbegrip en op de elementaire kansrekening. Ook zullen enkele begrippen en grootheden uit de wiskundige statistiek met de belangrijkste eigenschappen worden behandeld. De behandeling heeft uitsluitend een inleidend karakter en zal van wiskundig standpunt bezien niet altijd volledig bevredigend zijn.

Als aanvulling van dit college kan ten zeerste worden aanbevolen de bestudering van het boekje:

H. Freudenthal, "Waarschijnlijkheid en Statistiek".
Haarlem, De Erven F. Bohn N.V., 1957.

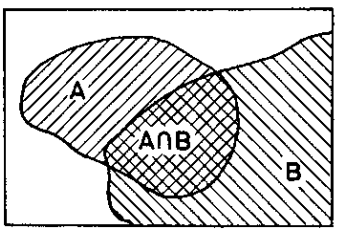
In duitse uitgave
"Wahrscheinlichkeit und Statistiek"
München, R. Oldenburg, 1963.

2. Gebeurtenissen

Uitspraken in de kansrekening hebben altijd betrekking op mogelijke gebeurtenissen. Bij het gooien met een dobbelsteen bestaat de uitkomstenruimte uit de getallen 1, 2, 3, 4, 5 en 6.
Gebeurtenissen zijn bv.:
de uitkomst van een worp is een 4;
de uitkomst van een worp is even;
de uitkomst van een worp is groter dan 3;
de uitkomst van een worp is oneven en kleiner dan 4;
de uitkomst van een worp is òf even, òf kleiner dan 3.

Bij het meten van het zwavelgehalte van een ruwe olie bestaat de uitkomstenruimte uit de getallen tussen 0 en 1.
Gebeurtenissen zijn bv.:
het gemeten zwavelgehalte ligt tussen 0.02 en 0.04;
het gemeten zwavelgehalte is niet groter dan 0.03;
het gemeten zwavelgehalte is groter dan 0.05.

Bij het schieten op een schietschijf (aannemend dat de schutter deze altijd raakt) bestaat de uitkomstenruimte uit de punten op de schietschijf.
Gebeurtenissen zijn bv.:
de kogel treft gebied A;
de kogel treft gebied A of gebied B;
de kogel treft gebied A en gebied B.



Uit deze en soortgelijke voorbeelden volgt dat voor wiskundige doeleinden een gebeurtenis als volgt gedefinieerd moet worden:
Een gebeurtenis is een deelverzameling van de uitkomstenruimte.
Bij elke gebeurtenis behoort een complementaire gebeurtenis $\bar{A}$ (niet A).
Bv. De gebeurtenis "de uitkomst van een worp met een dobbelsteen is even maar niet 4", heeft tot complementaire gebeurtenis: "de uitkomst van een worp met een dobbelsteen is oneven of 4".

Ga na wat de complementaire gebeurtenissen zijn van de gebeurtenissen in de voorafgaande voorbeelden!

Uiteraard is de gehele uitkomstenruimte een gebeurtenis, de zg. "zekere gebeurtenis", omdat zeker is dat de stochastische variabele een waarde uit deze "deel"-ruimte zal aannemen. De complementaire gebeurtenis daarvan is een gebeurtenis die nooit kan plaatsvinden: de zg.: "lege gebeurtenis". Voor rekendoeleinden is het echter zeer efficiënt ook deze lege gebeurtenis tot de verzameling van gebeurtenissen te tellen. Zij zal steeds met het teken $\emptyset$ worden aangeduid. De gehele uitkomstenruimte zullen we steeds met de letter U aanduiden.

Treden twee gebeurtenissen A en B gelijktijdig op, dan kunnen we dit als een nieuwe gebeurtenis C aanduiden. We gebruiken de notatie:

$$C = A \cap B \qquad (C = A \text{ en } B).$$

Bij het gooien met een dobbelsteen:
gebeurtenis A: $\underline{x} \in \{2, 4, 6\}$
gebeurtenis B: $\underline{x} \in \{1, 2, 3\}$
gebeurtenis A $\cap$ B: $\underline{x} \in \{2, 4, 6\} \cap \{1, 2, 3\} = \{2\}$.

Het is mogelijk dat twee gebeurtenissen A en B niet gelijktijdig kunnen optreden. Dan geldt:

$$A \cap B = \emptyset .$$

Gebeurtenissen welke niet gelijktijdig kunnen plaatshebben, noemen we "elkaar uitsluitende gebeurtenissen".

Soms is men er alleen in geïntersseerd of van twee genoemde gebeurtenissen A en B er minstens één voorkomt. Ook dit kunnen we als een nieuwe gebeurtenis C aanduiden. We gebruiken de notatie

$$C = A \cup B \qquad (C = A \text{ en/of } B).$$

Het is duidelijk dat voor elke gebeurtenis A geldt:

$$A \cup \overline{A} = U ,$$
$$A \cap \overline{A} = \emptyset .$$

De rekenregels voor de operaties $\cap$ en $\cup$ zijn analoog aan die voor getallen:

commutatieve eigenschappen
$$A \cap B = B \cap A$$
$$A \cup B = B \cup A ,$$

associatieve eigenschappen
$$A \cap (B \cap C) = (A \cap B) \cap C = A \cap B \cap C$$
$$A \cup (B \cup C) = (A \cup B) \cup C = A \cup B \cup C ,$$

distributieve eigenschappen
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

Voor de complementvorming gelden de regels:

$$\overline{A \cup B} = \overline{A} \cap \overline{B}$$
$$\overline{A \cap B} = \overline{A} \cup \overline{B} .$$

Hoewel de bewijzen voor deze regels heel eenvoudig zijn, zullen wij deze hier niet geven (probeer zelf). Wij zullen ze slechts in een enkel geval nodig hebben en ze dan zonder meer gebruiken.

We maken nu de volgende afspraak:
Als we kansuitspraken doen over de stochastische variabele $\underline{x}$ met uitkomstenruimte U, dan wordt eerst op ondubbelzinnige wijze een verzameling E van gebeurtenissen gedefinieerd. Deze verzameling E van gebeurtenissen moet de volgende eigenschappen bezitten:

1) als $A \in E$ dan is A een deelverzameling van U;
2) $U \in E$;
3) als $A \in E$ dan ook $\overline{A} \in E$;
4) als $A \in E$ en $B \in E$ dan ook $A \cup B \in E$;
5) als $A \in E$ en $B \in E$ dan ook $A \cap B \in E$.

Voor een wiskundig bevredigende opbouw van de kansrekening is het ook gewenst dat voldaan is aan de conditie:

4*) als $A_i \in E$, $i = 1, 2, \ldots\ldots$, dan ook $\bigcup_i A_i \in E$.

Van deze eigenschap zullen we echter geen expliciet gebruik maken.

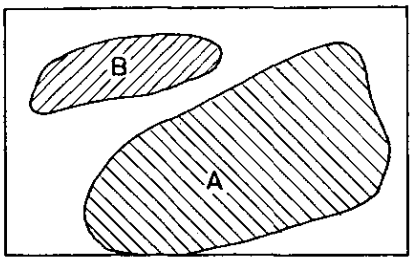
3. Kansverdelingen

Bij het intuitieve kansbegrip is "de kans op een bepaalde gebeurtenis" altijd een getal tussen 0 en 1 (of, als we in procenten werken, een getal tussen 0 en 100). Extreme gevallen zijn:
een gebeurtenis welke nooit plaats kan hebben, heeft de kans 0;
een gebeurtenis welke zeker plaats zal hebben, heeft de kans 1.

Bij de invoering van het wiskundige kansbegrip doen we hetzelfde.
Laat $\underline{x}$ een stochastische variabele zijn met uitkomstenruimte U. Het woord: stochastische variabele impliceert dat we beschikken over een mechanisme waarmee we op elk gewenst moment aan $\underline{x}$ een waarde x uit U kunnen toekennen. Zo'n mechanisme is bv. een dobbelsteen, een ton met loten, een roulette, een afschietmechanisme, de uitvoering van een meting volgens een bepaald voorschrift enz.. Het gebruik van dit mechanisme met het doel aan $\underline{x}$ een waarde toe te kennen, zullen we ook noemen "het realiseren van de stochastische variabele $\underline{x}$".
Als nu E de verzameling van gebeurtenissen is waarin we zijn geinteresseerd (en welke voldoet aan de eisen gesteld in de voorafgaande paragraaf) dan is aan het mechanisme een "kansverdeling" verbonden; d.w.z. door de structuur en de werking van het mechanisme is aan elke gebeurtenis A uit E een getal $P(A)$ toegekend, gelegen tussen 0 en 1. Dit getal is de kans dat, bij gebruik van het mechanisme, de stochastische variabele $\underline{x}$ een waarde aanneemt in de deelverzameling A van de uitkomstenruimte U.
Elk mechanisme heeft zijn eigen kansverdeling. Toch hebben al deze kansverdelingen enkele gemeenschappelijke eigenschappen.
Allereerst wordt aan de uitkomstenruimte U (dit is een gebeurtenis) de kans $P(U) = 1$ toegekend : U is de zekere gebeurtenis. Aan de lege gebeurtenis $\emptyset$ wordt de kans $P(\emptyset) = 0$ toegekend: $\emptyset$ is een nooit voorkomende gebeurtenis.
Een derde eigenschap wordt direct duidelijk uit het model van de schietschijf.
Laten we aannemen dat het afschiet mechanisme zo is geconstrueerd, dat de schijf altijd wordt geraakt, maar dat er verder geen enkele voorkeur bestaat voor het treffen van enig punt. (Intuitief is duidelijk wat dit betekent; op de wiskundige formulering daarvan komen we nog terug).

Een gebeurtenis is te identificeren met een deelgebied van de schijf. Bestaat er geen voorkeur voor enig punt, dan is de kans dat het gebied A getroffen wordt evenredig met het oppervlak van A zodat we kunnen stellen

$$P(A) = \frac{\text{Opp. A}}{\text{Opp schijf}} .$$



Laat B nu een tweede gebied op de schijf zijn, dat geen enkel punt met A gemeen heeft. Dus $A \cap B = \emptyset$. Deze twee gebieden representeren elkaar uitsluitende gebeurtenissen.
Nu is het duidelijk dat

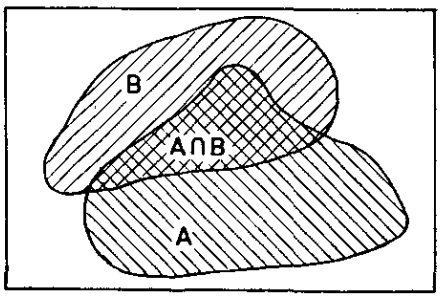
$$P(A \cup B) = \frac{\text{Opp}(A \cup B)}{\text{Opp schijf}}$$

$$= \frac{\text{Opp A + Opp B}}{\text{Opp schijf}}$$

$$= \frac{\text{Opp A}}{\text{Opp schijf}} + \frac{\text{Opp B}}{\text{Opp schijf}}$$

$$= P(A) + P(B)$$

Hebben de twee gebieden wel punten gemeen, sluiten de gebeurtenissen elkaar dus niet uit, dan geldt:



$$P(A \cup B) = \frac{\text{Opp}(A \cup B)}{\text{Opp schijf}}$$

$$= \frac{\text{Opp A + Opp B - Opp}(A \cap B)}{\text{Opp schijf}}$$

$$= \frac{\text{Opp A}}{\text{Opp schijf}} + \frac{\text{Opp B}}{\text{Opp schijf}} - \frac{\text{Opp}(A \cap B)}{\text{Opp schijf}}$$

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

Deze twee uitkomsten, die algemeen voor kansen gelden, worden samengevat in de "optelregel voor kansen":
Als A en B twee gebeurtenissen zijn in E dan geldt

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

Samenvattend kunnen we nu zeggen:
Een functie P, gedefinieerd op de verzameling E van gebeurtenissen, wordt een kansverdeling genoemd als zij voldoet aan de volgende eisen:

K1. Als $A \in E$ dan is $0 \leq P(A) \leq 1$

K2. $P(U) = 1$

K3. $P(\emptyset) = 0$

K4. Als $A \in E$ en $B \in E$ terwijl $A \cap B = \emptyset$
dan is $P(A \cup B) = P(A) + P(B)$

K5. Als $A \in E$ en $B \in E$
dan is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Van wiskundig standpunt bezien kan met minder eisen worden volstaan. Zo is de eis K5 overbodig en kan uit K4 worden afgeleid; we zullen dit echter niet aantonen. Wel zullen we hiervan gebruik maken als we van een gegeven functie P willen aantonen dat dit inderdaad een kansverdeling is.

Voor een wiskundig bevredigende opbouw van de kansrekening moet ook nog geëist worden:

K4* Als $A_i \in E$, $i = 1, 2, \ldots$, terwijl $A_i \cap A_j = \emptyset$ voor $i \neq j$
dan geldt:
$$P(\bigcup_i A_i) = \sum_i P(A_i).$$

Van deze eis zullen we bij onze elementaire inleiding echter geen expliciet gebruik maken.

$P(A)$ moet geïnterpreteerd worden als de "kans dat de stochastische variabele $\underline{x}$ bij realisering een waarde in A aanneemt". We zullen in plaats van $P(A)$ ook vaak de iets langere notatie $P(\underline{x} \in A)$ gebruiken.

4.1 Conditionele kansverdelingen

In dit hoofdstuk zijn we geinteresseerd in vragen van het volgende type:
wat is de kans dat de vraag naar een bepaald onderdeel uit een voorraadmagazijn in de komende week minstens even groot is als in deze week?
Wat is de kans dat een recruut met lengte tussen 170 en 175 cm meer weegt dan 160 pond?
Wat is de kans op een jongensgeboorte in een gezin met drie kinderen, welke alle drie meisjes zijn?
In al deze gevallen beschouwen we in feite paren stochastische variabelen $(\underline{x}, \underline{y})$:
de vraag $\underline{x}$ naar een bepaald onderdeel in deze week en de vraag $\underline{y}$ naar datzelfde onderdeel in de komende week;
de lengte $\underline{x}$ van een recruut en het gewicht $\underline{y}$ van diezelfde recruut;
in een gezin met 3 kinderen, allen van hetzelfde geslacht, waarin het vierde verwacht wordt: het geslacht $\underline{x}$ van de eerste drie kinderen en het geslacht $\underline{y}$ van het verwachte vierde kind.

Zo'n tweetal stochastische variabelen $(\underline{x}, \underline{y})$ is op zichzelf weer een stochastische variabele. Vergelijk bv. een punt op een schietschijf en de coördinaten van dat punt in een daarop aangebracht assenstelsel.
Het geven van een uitkomstenruimte W voor de stochastische variabele $\underline{z} = (\underline{x}, \underline{y})$ impliceert het geven van een uitkomstenruimte U voor de stochastische variabele $\underline{x}$ en een uitkomstenruimte V voor de stochastische variabele $\underline{y}$.
Ook impliceert een gebeurtenissenverzameling G voor $\underline{z} = (\underline{x}, \underline{y})$ een gebeurtenissenverzameling E voor $\underline{x}$ en een gebeurtenissenverzameling F voor $\underline{y}$.
Als nu $A \in E$ en $B \in F$ dan $(A, B) \in G$ dus dan is (A, B) een mogelijke gebeurtenis voor $\underline{z} = (\underline{x}, \underline{y})$.

De gebeurtenissen $A \in E$ voor $\underline{x}$ kunnen geidentificeerd worden met de gebeurtenissen (A, V) voor $\underline{z} = (\underline{x}, \underline{y})$, dwz: als we alleen op de stochastische variabele $\underline{x}$ letten, dan interessert het ons niet welke waarde $\underline{y}$ aanneemt. (Let erop dat V de zekere gebeurtenis voor $\underline{y}$ voorstelt!)

Evenzo worden de gebeurtenissen $B \in F$ voor $\underline{y}$ geïdentificeerd met de gebeurtenissen (U, B) voor $\underline{z} = (\underline{x}, \underline{y})$. (Let er weer op dat U de zekere gebeurtenis voor $\underline{x}$ voorstelt!)

Een kansverdeling voor $\underline{z} = (\underline{x}, \underline{y})$ leidt nu onmiddellijk naar een kansverdeling voor $\underline{x}$ en naar een kansverdeling voor $\underline{y}$ als we definiëren:

$$P(\underline{x} \in A) = P(\underline{x} \in A, \underline{y} \in V)$$
voor alle $A \in E$

en

$$P(\underline{y} \in B) = P(\underline{x} \in U, \underline{y} \in B)$$
voor alle $B \in F$.

Dat dit inderdaad kansverdelingen zijn, dus dat de aldus gedefinieerde functies P aan de eisen gesteld in hoofdstuk 3 voldoen blijkt als volgt:

1. $P(\underline{x} \in A)$ is gedefinieerd voor elke $A \in E$.
2. $0 \leq P(\underline{x} \in A) \leq 1$ omdat $0 \leq P(\underline{x} \in A, \underline{y} \in V) \leq 1$.
3. De zekere gebeurtenis voor $\underline{x}$ is U; (U, V) is echter de zekere gebeurtenis voor $(\underline{x}, \underline{y})$ dus:

$$P(\underline{x} \in U) = P(\underline{x} \in U, \underline{y} \in V) = 1.$$

4. Als $A_1 \in E$ en $A_2 \in E$ terwijl $A_1 \cap A_2 = \emptyset$ dan geldt:

$$\begin{aligned} P(\underline{x} \in A_1 \cup A_2) &= P(\underline{x} \in A_1 \cup A_2, \underline{y} \in V), \\ &= P((\underline{x} \in A_1, \underline{y} \in V) \text{ en } (\underline{x} \in A_2, \underline{y} \in V)) \\ &= P(\underline{x} \in A_1 \ \underline{y} \in V) + P(\underline{x} \in A_2, \underline{y} \in V) \\ &= P(\underline{x} \in A_1) + P(\underline{x} \in A_2) \end{aligned}$$

(bedenk dat uit $A_1 \cap A_2 = \emptyset$ volgt dat de gebeurtenissen $(\underline{x} \in A_1, \underline{y} \in V)$ en $(\underline{x} \in A_2, \underline{y} \in V)$ elkaar uitsluiten.

Bij gegeven kansverdeling voor de variabele $\underline{z} = (\underline{x}, \underline{y})$ wordt

$$P(\underline{x} \in A) = P(\underline{x} \in A, \underline{y} \in V)$$

de marginale kansverdeling voor de component $\underline{x}$ van $\underline{z}$ genoemd.

In voorgaande hebben we uit de kansverdeling voor $\underline{z} = (\underline{x}, \underline{y})$ een kansverdeling voor $\underline{x}$ afgeleid onder de voorwaarde dat we geen interesse hebben in de uitkomst van $\underline{y}$, dus onder voorwaarde dat $\underline{y} \in V$; dat $\underline{y}$ een zekere gebeurtenis is.

Nu willen we uit de kansverdeling voor $\underline{z} = (\underline{x}, \underline{y})$ een kansverdeling voor $\underline{x}$ afleiden onder de voorwaarde dat voor $\underline{y}$ een voorafgegeven gebeurtenis B plaats heeft. Het centrale probleem is hier feitelijk: levert de wetenschap dat voor $\underline{y}$ de gebeurtenis B plaats heeft ons informatie over het stochastische gedrag van $\underline{x}$?
We beschouwen in feite een nieuwe stochastische variabele

$$(\underline{x} \mid \underline{y} \in B).$$

Voor de vertikale streep staat de naam van de stochastische variabele, erachter staat de voorwaarde. Aan deze variabele gaan we nu enkele heuristische beschouwingen wijden.
De uitkomstenruimte voor $(\underline{x} \mid \underline{y} \in B)$ is weer U en de gebeurtenissenverzameling is ook weer E.
Als $A \in E$, dus als A een mogelijke gebeurtenis is voor $(\underline{x} \mid \underline{y} \in B)$, dan schrijven we inplaats van

$$(\underline{x} \mid \underline{y} \in B) \in A$$

ook

$$(\underline{x} \in A \mid \underline{y} \in B).$$

Het optreden van de gebeurtenis A voor de stochastische variabele $(\underline{x} \mid \underline{y} \in B)$ betekent voor $\underline{z} = (\underline{x}, \underline{y})$ het optreden van de gebeurtenis $(\underline{x} \in A, \underline{y} \in B)$. Het is dan ook redelijk te stellen dat de kans op het optreden van de gebeurtenis A, als we weten dat voor $\underline{y}$ de gebeurtenis B optreedt, evenredig (niet noodzakelijk gelijk!) is aan de kans op het optreden van de gebeurtenis (A, B) voor $\underline{z} = (\underline{x}, \underline{y})$.

D.w.z. we stellen

$$P(\underline{x} \in A \mid \underline{y} \in B) = \alpha \cdot P(\underline{z} \in (A,B)) = \alpha \cdot P(\underline{x} \in A, \underline{y} \in B)$$

waarin α een evenredigheidscoëfficient is welke niet van de variabele gebeurtenis A afhangt. Maar wellicht wel van de vaste gebeurtenis B !

Om α te bepalen zoeken we naar een gebeurtenis waarvan we de kans op voorkomen kennen. Dit is bv. de zekere gebeurtenis voor $(\underline{x} \mid \underline{y} \in B)$. Deze zekere gebeurtenis is uiteraard de gebeurtenis U. Nu geldt:

$$\begin{aligned} 1 &= P(\underline{x} \in U \mid \underline{y} \in B), \\ &= \alpha \cdot P(\underline{x} \in U, \underline{y} \in B), \\ &= \alpha \cdot P(\underline{y} \in B). \end{aligned}$$

(zie pagina 12)

Hieruit volgt:

$$\alpha = \frac{1}{P(\underline{y} \in B)} .$$

Tenminste als $P(\underline{y} \in B) \neq 0$. Dit laatste zullen we echter bij conditionele kansbeschouwingen steeds aannemen.
Deze heuristische beschouwingen suggereren dus dat de kans op de gebeurtenis A voor de stochastische variabele $\underline{x}$, als reeds bekend is dat voor $\underline{y}$ de gebeurtenis B optreedt, gesteld moet worden:

$$P(\underline{x} \in A \mid \underline{y} \in B) = \frac{P(\underline{x} \in A, \underline{y} \in B)}{P(\underline{y} \in B)} .$$

We moeten echter nog nagaan of de aldus gedefinieerde functie van A inderdaad aan de eisen voor kansverdelingen, als gespecificeerd in hoofdstuk 3, voldoet.

1. $P(\underline{x} \in A \mid \underline{y} \in B)$ is gedefinieerd voor elke $A \in E$.
2. $P(\underline{x} \in A \mid \underline{y} \in B) \geq 0$, duidelijk

3. $P(\underline{y} \in B) := P(\underline{x} \in U, \underline{y} \in B) \geq P(\underline{x} \in A, \underline{y} \in B)$
   omdat de gebeurtenis $(\underline{x} \in U, \underline{y} \in B)$ de gebeurtenis $(\underline{x} \in A, \underline{y} \in B)$ impliceert.
   Hieruit volgt dat $P(\underline{x} \in A \mid \underline{y} \in B) \leq 1$ is.
4. De zekere gebeurtenis $(\underline{x} \in U \mid \underline{y} \in B)$ heeft kans 1 wegens de speciale keuze van α.
5. Als $A_1 \in E$ en $A_2 \in E$, terwijl $A_1 \cap A_2 = \emptyset$ dan geldt:

$$P(\underline{x} \in A_1 \cup A_2 \mid \underline{y} \in B) = P(\underline{x} \in A_1 \cup A_2, \underline{y} \in B) / P(\underline{y} \in B)$$
$$= P\Big((\underline{x} \in A_1, \underline{y} \in B) \text{ en } (\underline{x} \in A_2 \text{ en } \underline{y} \in B)\Big) / P(\underline{y} \in B).$$

Tussen de grote haken staan nu, wegens $A_1 \cap A_2 = \emptyset$, twee elkaar uitsluitende gebeurtenissen $(\underline{x} \in A_1, \underline{y} \in B)$ en $(\underline{x} \in A_2, \underline{y} \in B)$.
Dus

$$P(\underline{x} \in A_1 \cup A_2 \mid \underline{y} \in B) = \{P(\underline{x} \in A_1, \underline{y} \in B) + P(\underline{x} \in A_2, \underline{y} \in B)\} / P(\underline{y} \in B)$$
$$= \frac{P(\underline{x} \in A_1, \underline{y} \in B)}{P(\underline{y} \in B)} + \frac{P(\underline{x} \in A_2, \underline{y} \in B)}{P(\underline{y} \in B)}$$
$$= P(\underline{x} \in A_1 \mid \underline{y} \in B) + P(\underline{x} \in A_2 \mid \underline{y} \in B).$$

Aan de kanseisen K1, K2, K3 en K4 is dus voldaan.
Na deze heuristische beschouwingen definieren we de conditionele kans voor $\underline{x}$ op de gebeurtenis A onder voorwaarde dat voor $\underline{y}$ de gebeurtenis B optreedt, $P(\underline{y} \in B) \neq 0$ veronderstellend, door

$$P(\underline{x} \in A \mid \underline{y} \in B) = \frac{P(\underline{x} \in A, \underline{y} \in B)}{P(\underline{y} \in B)} \quad \text{voor elke } A \in E.$$

We kunnen ook schrijven:

$$P(\underline{x} \in A, \underline{y} \in B) = P(\underline{x} \in A \mid \underline{y} \in B) \cdot P(\underline{y} \in B)$$

welke formule ook geldt als $P(\underline{y} \in B) = 0$.

4.2 Onafhankelijke stochastische variabelen

Uiteraard kan het voorkomen dat paren stochastische variabelen bekeken worden, waarbij informatie over de ene variabele geen informatie over de kansverdeling van de andere variabele verschaft. Het tweetal variabelen wordt dan stochastisch onafhankelijk genoemd. De conditionele kansverdeling van de variabele $\underline{x}$, gegeven dat voor de variabele $\underline{y}$ de gebeurtenis B optreedt is nu niet afhankelijk van deze laatste informatie, wat B ook is. De informatie dat voor $\underline{y}$ de gebeurtenis B zal optreden mogen we daarom bv. vervangen door de informatie dat voor $\underline{y}$ de zekere gebeurtenis V zal optreden. Dus:

$$P(\underline{x} \in A \mid \underline{y} \in B) = P(\underline{x} \in A \mid \underline{y} \in V) \text{ voor elke } B \in V$$

$$= \frac{P(\underline{x} \in A, \underline{y} \in V)}{P(\underline{y} \in V)} = P(\underline{x} \in A)$$

De laatste overgang volgt uit $P(\underline{x} \in A, \underline{y} \in V) = P(\underline{x} \in A)$ en $P(\underline{y} \in V) = 1$.

Hieruit volgt: als de stochastische variabelen $\underline{x}$ en $\underline{y}$ onafhankelijk zijn, dan geldt voor elke gebeurtenis A van $\underline{x}$ en elke gebeurtenis B van $\underline{y}$ dat

$$P(\underline{x} \in A \mid \underline{y} \in B) = P(\underline{x} \in A) ,$$

$$P(\underline{y} \in B \mid \underline{x} \in A) = P(\underline{y} \in B) ,$$

$$P(\underline{x} \in A , \underline{y} \in B) = P(\underline{x} \in A).\ P(\underline{y} \in B).$$

Deze laatste formule, welke geen uitspraken meer bevat over conditionele waarschijnlijkheden, wordt ook vaak gebruikt als definitie voor onafhankelijkheid van de twee stochastische variabelen $\underline{x}$ en $\underline{y}$. Uit de formule voor de conditionele kansverdeling en uit deze laatste formule volgt dan direct dat

$$P(\underline{x} \in A \mid \underline{y} \in B) = P(\underline{x} \in A)$$

voor elke $A \in E$ en elke $B \in F$.

Ook voor meer dan twee stochastische variabelen kan het begrip "onafhankelijk" gedefinieerd worden.

Beschouw het n-tal stochastische variabelen $\underline{x}_i$, i = 1, ..., n met respectievelijk de uitkomstenruimten U_i en verzameling E_i van interessante gebeurtenissen.

Kiezen we hieruit, voor $m \leq n$, een m-tal variabelen $\underline{x}_{i_1}, \ldots, \underline{x}_{i_m}$ dan vormt de vector $(\underline{x}_{i_1}, \ldots, \underline{x}_{i_m})$ een nieuwe stochastische variabele met uitkomstenruimte $(U_{i_1}, \ldots, U_{i_m})$ en verzameling van interessante gebeurtenissen $(E_{i_1}, \ldots, E_{i_m})$.

Voorbeelden zijn: $(\underline{x}_1, \underline{x}_2)$, $(\underline{x}_1, \underline{x}_2, \underline{x}_3)$, ... , $(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_{n-1})$.

De stochastische variabelen $\underline{x}_i$, i = 1 , ..., n, worden stochastisch onafhankelijk genoemd als voor elke i geldt dat $\underline{x}_i$ stochastisch onafhankelijk is van de op deze wijze uit de overige n - 1 variabele gevormde nieuwe stochastische variabelen.

Onafhankelijkheid van het drietal stochastische variabelen $\underline{x}_1$, $\underline{x}_2$ en $\underline{x}_3$ betekent bv. dat de stochastische variabelen

$\underline{x}_1$, $\underline{x}_2$
$\underline{x}_1$, $\underline{x}_3$
$\underline{x}_1$, $(\underline{x}_2, \underline{x}_3)$
$\underline{x}_2$, $\underline{x}_3$
$\underline{x}_2$, $(\underline{x}_1, \underline{x}_3)$
$\underline{x}_3$, $(\underline{x}_1, \underline{x}_2)$

stochastisch onafhankelijk van elkaar moeten zijn.

Als de stochastische variabelen $\underline{x}_i$, i = 1, ..., n onafhankelijk van elkaar zijn dan geldt voor elke $A_i \in E_i$

$$P(\underline{x}_1 \in A_1, \underline{x}_2 \in A_2, \ldots, \underline{x}_n \in A_n) = P(\underline{x}_1 \in A_1) \cdot P(\underline{x}_2 \in A_2) \cdot \ldots P(\underline{x}_n \in A_n)$$

Dit is de zg. vermenigvuldigingsregel voor stochastisch onafhankelijke variabelen.

(1) $A_i \cap A_j = \emptyset$ als $i \neq j$ is; de gebeurtenissen sluiten elkaar dus uit,

(2) $\bigcup_i A_i = U$; de gebeurtenissen vormen een zg. compleet stelsel.

De condities (1) en (2) drukken uit dat bij realisering van $\underline{x}$ één en niet meer dan één van de gebeurtenissen A_i , $i = 1.2 \ldots\ldots$, zal optreden. Voor elke gebeurtenis $A \in E$ geldt nu:

$$A = A \cap U$$

$$= A \cap (A_1 \cup A_2 \cup A_3 \ldots\ldots) \qquad \text{(wegens (2))}$$

$$= (A \cap A_1) \cup (A \cap A_2) \cup \ldots\ldots \qquad \text{(distributieve eigenschap voor gebeurtenissen.)}$$

Het is duidelijk dat ook geldt:

$$(A \cap A_i) \cap (A \cap A_j) = \emptyset . \qquad \text{(wegens (1))}$$

Uit de regels K4 of K4* van de kansrekening (zie pag. 10) volgt nu direct:

$$P(A) = \sum_i P(A \cap A_i)$$

dus, wegens $P(A \cap A_i) = P(A \mid A_i) . P(A_i)$ (zie pag. 15).

$$P(A) = \sum_i P(A \mid A_i) . P(A_i) \qquad \text{voor elke } A \in E.$$

Deze relatie tussen de kans op voorkomen van de gebeurtenis A en de kans op voorkomen van de conditionele gebeurtenissen $(A \mid A_i)$ wordt in de literatuur de "law of total probability" genoemd.

In vele toepassingen blijkt het bij gegeven gebeurtenis A mogelijk de gebeurtenissen A_i , $i = 1, 2 \ldots$, zo te kiezen dat zowel $P(A_i)$ als $P(A \mid A_i)$ direct bekend of gemakkelijk te berekenen zijn. Het vormen van de gebeurtenissen $(A \mid A_i)$ noemt men het conditioneren van de gebeurtenis A naar de gebeurtenissen $A_1, A_2, \ldots\ldots$.

5. Discrete stochastieken

In dit en in enkele volgende hoofdstukken zullen we nader ingaan op enkele veel voorkomende kansverdelingen. Tenzij anders vermeld zullen we ons beperken (zonder dit steeds te herhalen) tot stochastische variabelen $\underline{x}$, welke de verzameling van de reële getallen of een deel daarvan tot uitkomstenruimte hebben.

Een stochastische variabele $\underline{x}$ met een eindige of aftelbaar oneindige uitkomstenruimte wordt een discrete stochastiek genoemd.
Een veel voorkomende discrete stochastiek is de zg. bernoulli stochastiek.
De uitkomstenruimte U bevat slechts twee elementen welke te identificeren zijn met de getallen 0 en 1; dus U = (0,1):

$$P(\underline{x} = 1) = p \quad ,$$
$$P(\underline{x} = 0) = 1 - p = q.$$

Deze bernoulli stochastiek treft men o.a. aan bij het kruis en munt spel. Een kruisworp wordt geidentificeerd met de gebeurtenis (1), een muntworp met de gebeurtenis (0). Verder is, bij een eerlijk kruis en munt spel:

$$P(\underline{x} = 1) = P(\underline{x} = 0) = 1/2.$$

Bij een productieproces van discrete elementen, bv. transistors, gloeilampen etc., is het doorgaans niet te vermijden dat ook defecte elementen geproduceerd worden. Bij een stabiel proces is de kans dat een geproduceerd element defect is een constante gelijk aan p. De kans dat zo'n element niet defect is, is dan uiteraard 1 - p. De toestand van een geproduceerd element is dus een bernoulli stochastiek.

Bij de geboorte van kinderen is het geslacht van het kind dat geboren gaat worden een bernoulli stochastiek met:

$$P(\text{het kind is een jongen}) = 0.514$$
$$P(\text{het kind is een meisje}) = 0.486.$$

Een discrete stochastiek met een uit meer dan twee elementen bestaande, doch eindige uitkomstenruimte U treft men o.a. aan bij het gooien met een dobbelsteen. Dan is U = (1, 2, ..., 6).
Er zijn in het totaal $2^6 = 64$ verschillende deelverzamelingen van U aan te wijzen, d.w.z. bij het gooien met een dobbelsteen kunnen we spreken over 2^6 verschillende gebeurtenissen. Het is echter niet nodig de kans op voorkomen van elk van deze 64 gebeurtenissen afzonderlijk op te geven. De regels voor het combineren van gebeurtenissen zijn zo dat al deze 64 gebeurtenissen gegenereerd kunnen worden uit de 6 "elementaire", elkaar uitsluitende gebeurtenissen (1), (2), (3), (4), (5) en (6). De regels van de kansrekening op hun beurt zijn zo dat uit de kansen op deze elementaire gebeurtenissen de kans op elke willekeurige andere gebeurtenis kan worden berekend.

Voorgaande eigenschap geldt algemeen voor discrete stochastische variabelen, dus ook als U een aftelbaar oneindig aantal elementen bevat. Bovendien mag in de gevallen die wij zullen beschouwen steeds verondersteld worden dat deze elementaire gebeurtenissen bestaan uit de (discrete) elementen van de uitkomstenruimte zelf.
De elementen van de uitkomstenruimte U van een discrete stochastische variabele $\underline{x}$ zijn steeds te identificeren met de gehele getallen.
Bij een eindige uitkomstenruimte met n elementen kunnen we dus aannemen dat U = (1, 2, ..., n) en bij een aftelbaar oneindige uitkomstenruimte dat U = (1, 2, ...), of dat U = (....., - 2 - 1, 0, 1, 2, ...).
Is $\underline{x}$ een element van U, dan zullen we voor de elementaire gebeurtenis ($\underline{x} \in (x)$) de kortere notatie ($\underline{x} = x$) gebruiken.

De functie $f(x) = P(\underline{x} = x)$ gedefinieerd op U wordt de "frequentieverdeling" van $\underline{x}$ genoemd.

Aangezien minstens één van de gebeurtenissen ($\underline{x} = x$), $x \in U$ plaats heeft vormen de elementaire gebeurtenissen tesamen de zekere gebeurtenis U zodat geldt (regel K4, pag. 10):

$$\sum_{x \in U} P(\underline{x} = x) = 1.$$

Uit deze laatste relatie volgt:
is U aftelbaar oneindig en van de vorm U = (1, 2, ...) dan is

$$\lim_{x \to \infty} f(x) = 0;$$

is U aftelbaar oneindig en van de vorm U = (.., - 2, -1, 0, 1, 2, ...) dan is

$$\lim_{x \to -\infty} f(x) = \lim_{x \to +\infty} f(x) = 0.$$

Naast de frequentieverdeling f(x) kennen we ook de cumulatieve distributiefunctie F(x):

$$F(x) = P(\underline{x} \leq x) \quad , x \in U \quad .$$

Kennelijk geldt:

$$F(x) = \sum_{i \leq x} P(\underline{x} = i) = \sum_{i \leq x} f(i).$$

(Regel K4, pag. 10 en definitie van f(i)).

Omgekeerd geldt ook:

Als U = (1, 2, ..., n) of U = (1, 2, ...) dan is

$$f(1) = F(1)$$
$$f(x) = F(x) - F(x - 1), \; x > 1$$

en als U = (....., - 2, - 1, 0, 1, 2, ...)

$$f(x) = F(x) - F(x - 1)$$

met randvoorwaarde

$$\lim_{x \to -\infty} f(x) = 0.$$

Hieruit volgt:

1. Het stochastische gedrag van een discrete stochastische variabele $\underline{x}$ is volledig bepaald zowel door de frequentieverdeling $f(x)$ als door de cumulatieve distributiefunctie $F(x)$;
2. De cumulatieve distributiefunctie is een monotoon niet-dalende functie van x;
3. Als U eindig is, dus als $U = (1, 2, \ldots, n)$ dan geldt $F(n) = 1$. Als U aftelbaar, oneindig is en van de vorm $U = (1, 2, \ldots)$ dan geldt:

$$\lim_{x \to \infty} F(x) = 1.$$

Is U van de vorm $(\ldots - 2, -1, 0, 1, 2, \ldots)$ dan geldt bovendien:

$$\lim_{x \to -\infty} F(x) = 0.$$

Voorbeelden van discrete verdelingen

a. Discrete uniforme verdelingen. Bij een eerlijk kruis en munt spel zijn de kansen op de elementaire gebeurtenissen (kruis) en (munt) gelijk aan 1/2:

$$P(\underline{x} = \text{kruis}) = P(\underline{x} = \text{munt}) = 1/2.$$

Bij een eerlijke dobbelsteen zijn de kansen op de elementaire gebeurtenissen (1), (2), (3), (4), (5), (6) gelijk aan 1/6

$$P(x = i) = 1/6, \quad ,i = 1, \ldots, 6$$

De totale kans 1 is dus "uniform verdeeld" over de elementaire gebeurtenissen. Algemener: men zegt dat een discrete stochastische grootheid $\underline{x}$ met eindige uitkomstenruimte $U = (1, 2, \ldots n)$ een uniforme verdeling bezit als:

$$f(x) = P(\underline{x} = x) = 1/n \quad \text{voor alle } x \in U.$$

Alle elementaire gebeurtenissen (1), (2), ..., (n) hebben gelijke kans op voorkomen.

b. De binomiale verdeling. Als voorbeeld van een bernoulli stochastiek hebben we genomen de kans op productie van een defecte gloeilamp. Deze kans zegt iets over de kwaliteit van het productieproces. De afnemer zal echter minder geinteresseerd zijn in deze kans p dan in het aantal gloeilampen dat defect is in een partij van n stuks welke hij wil kopen. Dat aantal is uiteraard weer een stochastische variabele die afhangt van p en n.

Eenzelfde soort stochastische variabele treffen aan bij het kruis en munt spel als we voor elke keer dat kruis gegooid wordt een gulden uitbetaald krijgen en als munt gegooid wordt niets ontvangen. Wij zijn dan geinteresseerd in de opbrengst van het spel na precies n keer spelen.
Beschouw de n bernoulli stochastieken $\underline{x}_i$, i = 1, ..., n met $P(\underline{x}_i = 1) = p$ en $P(\underline{x}_i = 0) = 1 - p = q$, onafhankelijk van i. Bovendien nemen we aan dat de n variabelen $\underline{x}_i$ onderling onafhankelijk zijn. We construeren dan een nieuwe stochastische variabele*

$$\underline{y} = \sum_{i=1}^{n} \underline{x}_i$$

met uitkomstenruimte U = (0, 1, 2, ..., n).
$\underline{y}$ wordt een binomiale stochastiek genoemd en de kansverdeling van $\underline{y}$ heet de binomiale kansverdeling.
Volgens voorgaande is deze binomiale kansverdeling volledig bepaald door de binomiale frequentieverdeling welke we nu zullen bepalen.

De stochastische variabele $\underline{y}$ neemt dan en slechts dan de waarde m aan (m is één van de getallen 0, 1, 2, 3, ..., n) als precies m van de variabele $\underline{x}_i$ de waarde 1 en de andere m - n variabelen $\underline{x}_i$ de waarde 0 aannemen.
Uit de onafhankelijkheid van de variabelen $\underline{x}_i$ volgt P(gebeurtenis met m bepaalde variabelen $\underline{x}_i$ gelijk 1 en de n - m andere variabelen $\underline{x}_i$ gelijk 0) = $p^m(1-p)^{n-m}$, een uitkomst die niet afhangt van welke variabelen de waarde 1 en welke de waarde 0 aannemen.

---

* Onder de som van twee stochastische variabelen $\underline{x}_1$ en $\underline{x}_2$ (notatie $\underline{x}_1 + \underline{x}_2$) verstaat men een nieuwe stochastische variabele $\underline{y}$ die gerealiseerd wordt door een realisering x_1 van $\underline{x}_1$ op te stellen bij een realisering x_2 van $\underline{x}_2$.

Verder geldt:

$$P(\underline{y} = m) = P\left(\sum_{i=1}^{n} \underline{x}_i = m\right)$$

$$= P\Big((\underline{x}_1 = 1 , \underline{x}_2 = 1 , \ldots\ldots , \underline{x}_m = 1 , \underline{x}_{m+1} = 0, \ldots\ldots, \underline{x}_n = 0)$$

$$\text{of} (\underline{x}_1 = 0, \underline{x}_2 = 1 \quad , \ldots\ldots, \underline{x}_m = 1, \underline{x}_{m+1} = 1, \ldots, \underline{x}_n = 0)$$

$$\text{of} \ldots\ldots$$

$$\text{of} \ (\underline{x}_1 = 0 , \ldots \underline{x}_{n-m} = 0, \underline{x}_{n-m+1} = 1 , \ldots \ldots, \underline{x}_n = 1)\Big)$$

De gebeurtenissen tussen ( ) sluiten elkaar uit, dus:

$$P(\underline{y} = m) = P(\underline{x}_1 = 1 , \underline{x}_2 = 1 , \ldots\ldots , \underline{x}_m = 1 , \underline{x}_{m+1} = 0, \ldots\ldots , \underline{x}_n = 0)$$

$$+ P(\underline{x}_1 = 0 , \underline{x}_2 = 1 , \ldots\ldots , \underline{x}_m = 1 , \underline{x}_{m+1} = 1 , \ldots\ldots , \underline{x}_n = 0)$$

$$+$$

$$+ P(\underline{x}_1 = 0 , \ldots\ldots , \underline{x}_{n-m} = 0 , \underline{x}_{n-m+1} = 1 , \ldots\ldots , \underline{x}_n = 1)$$

Volgens bovenstaande redenering zijn echter alle kansen in het rechterlid gelijk aan $p^m(1 - p)^{n-m}$. Blijft nog over te bepalen hoeveel termen in het rechterlid staan. Dit aantal is gelijk aan het aantal verschillende mogelijkheden waarin we aan m variabelen $\underline{x}_i$ de waarde 1 en aan de andere n - m de waarde 0 kunnen toekennen. Dit aantal is gelijk aan :

$$\binom{n}{m} = \frac{n\ !}{m\ !\ (n-m)\ !} \quad ,$$

waaruit volgt voor de binomiale frequentieverdeling:

$$f(m) = P(\underline{y} = m) = \frac{n\ !}{m\ !\ (n-m)\ !}\ p^m(1 - p)^{n-m} \quad , \ m = 0, 1 , \ldots\ldots, n.$$

De getallen $\frac{n\,!}{m\,!\,(n-m)\,!}$ zijn dezelfde als voorkomend bij het binomium van Newton (de binomiaal formule, zie Wiskunde I, hoofdstuk III, 2).
De cumulatieve binomiale distributie formule is:

$$F(m) = P[y \leq m] = \sum_{i=0}^{m} \frac{n\,!}{i\,!(n-i)\,!}\, p^i(1-p)^{n-i},\quad m = 0, 1, \ldots, n.$$

(Toon zelf aan dat $F(n) = 1$).

De binomiale frekwentieverdeling en de cumulatieve binomiale distributiefunctie zijn voor de practisch voorkomende waarden van p en n in de meeste statistische hand- en tabellenboeken getabelleerd.

c. De Poisson stochastiek. Een poissonverdeelde stochastische variabele $\underline{x}$ bezit een aftelbaar oneindige uitkomstenruimte, bestaande uit de niet-negatieve gehele getallen: $U = \{0, 1, 2, ...\}$. De verzameling van interessante gebeurtenissen wordt gegenereerd door de elementaire gebeurtenissen (0), (1), (2), ...
De frekwentieverdeling van $\underline{x}$ heeft de vorm

$$f(x) = P[\underline{x} = x] = e^{-\lambda} \frac{\lambda^x}{x!} \quad , \; x = 0, 1, 2, ...$$

De cumulatieve distributiefunctie van $\underline{x}$ is:

$$F(x) = P[\underline{x} \leq x] = e^{-\lambda} \sum_{i=0}^{x} \frac{\lambda^i}{i!} \quad , \; x = 0, 1, 2, ...$$

De parameter λ is positief: de betekenis van deze parameter zal in hoofdstuk 9 duidelijk worden.

Poissonverdeelde stochastische variabelen komen veelvuldig voor bij wachttijd problemen. Zo is het gebruikelijk bij beschouwingen over bezetting van telefoonlijnen te veronderstellen dat het aantal gespreksaanvragen per tijdseenheid een poissonverdeling bezit.

De poissondistributie is getabelleerd in de meeste statistische handboeken en tabellenboeken.

6. Continue stochastieken.
Behoort bij de stochastische variabele $\underline{x}$ een een-dimensionaal continuum U als uitkomstenruimte (bv. de verzameling van de reële getallen of van de niet-negatieve getallen, het interval (0,1) enz.) dan zijn we doorgaans geïnteresseerd in gebeurtenissen welke door deelintervallen van U of combinaties daarvan worden gerepresenteerd.
Bij slijtageproeven met autobanden van een bepaalde samenstelling uitgevoerd volgens een welomschreven testvoorschrift, kunnen we vragen naar de volgende gebeurtenissen:
Ligt de gemeten slijtage tussen 20 en 30 gram per 1000 km rijafstand?

Zo ja, dan voldoet hij aan een gestelde kwaliteitseis.
Is de gemeten slijtage per 1000 km groter dan 35 gram of lager dan 28 gram;
zo ja, dan is aan een gestelde kwaliteitseis niet voldaan.

Wij kunnen bij onze beschouwingen steeds aannemen dat U bestaat uit de complete getallenrechte. Gedeelten daarvan welke nooit zullen voorkomen (negatieve slijtages bv.) geven we een kans 0 op voorkomen. Van kanstheoretisch standpunt uit bezien genereren de (open, gesloten en half open) intervallen inderdaad een verzameling E van gebeurtenissen welke aan de in hoofdstuk 2 gestelde eisen voldoet. De kans op een willekeurige gebeurtenis uit E is dus te berekenen als de kansen $P(\underline{x} \in I)$ voor alle intervallen $I \in U$ bekend zijn. Deze functie $P[\underline{x} \in I]$, gedefinieerd op de verzameling van de deelintervallen van de reële getallenrechte U is een zg. "intervalfunctie". Zulke functies zijn in het algemeen moeilijk hanteerbaar. Men kan echter aantonen dat het stochastische karakter van de variabele $\underline{x}$, evenals in het discrete geval, volledig beschreven kan worden door de distributiefunctie

$$F(x) = P[\underline{x} \leqslant x] \quad , \text{ gedefinieerd voor alle } x \in U.$$

Als de kansverdeling van $\underline{x}$ gegeven is door de intervalfunctie $P[\underline{x} \in I]$ voor alle intervallen I van U dan is ook de distributiefunctie $F(x)$ bekend. Men kan aantonen dat $F(x)$ de volgende eigenschappen bezit (het bewijs wordt hier niet gegeven!):

F1. $0 \leqslant F(x) \leqslant 1$,

F2. $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$,

F3. $F(x)$ is een monotone, niet dalende functie van x,

F4. $F(x)$ is voor elke x continu naar rechts, d.w.z.
$\lim_{\delta \to +0} F(x + \delta) = F(x)$ voor elke $x \in U$

Omgekeerd kan men aantonen dat elke functie $F(x)$ welke aan deze eisen voldoet als distributiefunctie kan optreden.

Van nu af aan zullen we, tenzij anders vermeld, het stochastisch karakter van een stochastische variabele steeds beschrijven door een distributiefunctie.

Een stochastische variabele $\underline{x}$ wordt continu genoemd als de bijbehorende distributiefunctie continu is voor elke $x \in U(-\infty, \infty)$ en, hoogstens met uitzondering van een eindig aantal punten, differentieerbaar is over het uitkomstengebied $U(-\infty < x < \infty)$. In vele gevallen is het in aanmerking te nemen uitkomstengebied slechts een echt deel van de getallenrechten $(-\infty, \infty)$. (zie de voorbeelden a, b, en c aan het einde van dit hoofdstuk). Ter voorkoming van ingewikkelde redeneringen en formuleringen bespreken we in het volgende uitsluitend stochastische variabelen die over hun gehele uitkomstengebied differentieerbaar zijn. Aan de "randen" moeten we dan spreken van rechts resp. links differentieerbaarheid.

Voor een continue stochastische variabele $\underline{x}$ geldt $P[\underline{x} = x] = 0$ voor elke $x \in U$: dus zelfs als $\underline{x} = x$ een mogelijke gebeurtenis is voor de stochastische variabele $\underline{x}$, dan is toch de kans op deze gebeurtenis gelijk 0.
Het bewijs volgt uit:

$$P[\underline{x} = x] = \lim_{\delta \to 0}(F(x + \delta) - F(x)) = 0$$

De afgeleide $f(x) = F'(x)$ van de distributiefunctie $F(x)$ van een continue stochastische variabele $\underline{x}$ wordt de"kansdichtheidsfunctie" van $\underline{x}$ genoemd. Ter verduidelijking van deze term schrijven we

$$f(x) = \lim_{\delta \to 0} \frac{F(x + \delta) - F(x)}{\delta}$$

$$= \lim_{\delta \to 0} \frac{P(x < \underline{x} \leq x + \delta)}{\delta}$$

De uitdrukking $\frac{P(x < \underline{x} \leq x + \delta)}{\delta}$ stelt de "gemiddelde kans" voor per eenheid van lengte welke aan het interval $(x < \underline{x} \leq x + \delta)$ is toegekend of populair gezegd: zij geeft aan hoe dicht de kans $P[x < \underline{x} \leq x + \delta]$ gespreid ligt over dit interval als we deze kans gelijkmatig over dit interval uitgesmeerd denken. In het limietgeval $\delta \longrightarrow 0$ kunnen we dus spreken van de kansdichtheid in het punt x.

Uit de monotonie van $F(x)$ (zie eigenschap F3) volgt:

$$f(x) \geqslant 0 \qquad \text{voor elke } x;$$

de kansdichtheid van $\underline{x}$ is dus nooit negatief.

Uit de een bekende eigenschap van de integraalrekening volgt:

1. $P[a \leqslant \underline{x} \leqslant b] = \int_a^b f(x)\, d\, x$

2. $F(x) = \int_{-\infty}^{x} f(u)\, d\, u$

3. $F(\infty) = \int_{-\infty}^{\infty} f(x)\, d\, x = 1$

Omgekeerd geldt: elke functie $f(x)$ met de eigenschappen

$$f(x) \geqslant 0 \qquad \text{voor elke } x$$

en $$\int_{-\infty}^{\infty} f(x)\, d\, x = 1,$$

kan optreden als kansdichtheidsfunctie.
De functie $F(x) = \int_{-\infty}^{x} f(u)\, du$ voldoet dan aan de bovengenoemde
eisen voor een distributiefunctie.
Veel voorkomende distributiefuncties voor continue stochastische variabelen zijn:

a. <u>de uniforme distributie</u>

stochastische variabele $\underline{x}$; uitkomstenruimte het interval (a, b)
distributiefunctie:

$$f(x) = \frac{x-a}{b-a} \qquad a \leqslant x \leqslant b;$$

dichtheidsfunctie:

$$f(x) = \frac{1}{b-a} \qquad a \leqslant x \leqslant b.$$

b. De negatief exponentiële distributie

stochastische variabele $\underline{x}$; uitkomstenruimte het interval $(0 \quad \infty)$; distributiefunctie:

$$F(x) = 1 - e^{-\lambda x} \qquad x \geq 0$$

dichtheidsfunctie:

$$f(x) = \lambda e^{-\lambda x} \qquad x \geq 0.$$

waarin λ een vast getal > 0 is.

De negatief exponentiële distributie treedt vaak op in combinatie met de poissonverdeling. Zo volgt het aantal radioactieve atomen dat in het tijdsverloop t uit elkaar valt een poissonverdeling:

$$P[\underline{x} = n] = e^{-\lambda t} \frac{(\lambda t)^n}{n!}.$$

Het tijdsverloop t waarin geen atomen uit elkaar vallen, is een negatief exponentieel verdeelde stochastische variabele met kansdichtheid

$$f(t) = P[\underline{t} = t] = \lambda e^{-\lambda t} \qquad t \geq 0 \quad .$$

c. De normale verdeling

Ook genoemd de gaussverdeling, de De Moivreverdeling; stochastische variabele $\underline{x}$, uitkomstenruimte de reële getallenrechte;
distributiefunctie:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\frac{(t-\mu)^2}{\sigma^2}} \, dt = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp - \frac{1}{2} \frac{(t-\mu)^2}{\sigma^2} \, dt$$

dichtheidsfunctie

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

Hoewel de dichtheids- en de distributiefunctie van de normale verdeling een gecompliceerde vorm hebben, speelt deze toch een overheersende rol in de statistiek.

De normale verdeling bevat twee parameters welke we in een volgend hoofdstuk nader zullen bespreken.

7. Functies van stochastische variabelen

Zij $f(x)$ een functie gedefinieerd op de uitkomstenruimte U van de stochastische variabele $\underline{x}$. Dan is $\underline{y} = f(\underline{x})$ een nieuwe stochastische variabele zodat als $\underline{x}$ bij realisering de waarde $x \in U$ aanneemt, tegelijkertijd $\underline{y}$ de waarde $y = f(x)$ aanneemt. Spelen we bv. kruis en munt en spreken we af dat we 5 gulden ontvangen als kruis geworpen wordt en dat we 3 gulden moeten betalen bij een muntworp, dat wordt het spel beschreven door een stochastische variabele $\underline{x}$ met uitkomstenruimte (kruis, munt), terwijl de opbrengst van het spel beschreven wordt door een stochastische variabele $\underline{y}$ met uitkomstenruimte $(5, -3)$ en wel zo dat:

$\underline{y} = 5$ als $\underline{x}$ = kruis
$y = -3$ als $\underline{x}$ = munt.

Onder zeer ruime voorwaarden voor de functie $y = f(x)$ geldt dat met elke gebeurtenis betreffende $\underline{y}$ een gebeurtenis betreffende $\underline{x}$ correspondeert en omgekeerd. Bij kansbeschouwingen worden uiteraard aan deze corresponderende gebeurtenissen dezelfde kansen toegevoegd. Zij $\underline{x}$ een stochastische variabele met de reële getallenrechte tot uitkomstenruimte U. Zij bovendien $y = f(x)$ een reële functie gedefinieerd op de reële getallenrechte U. De stochastische variabele $\underline{y} = f(\underline{x})$ heeft dan weer deze reële getallenrechte U tot uitkomstenruimte. Willen we de distributiefunctie $G(y)$ van $\underline{y}$ bepalen dan moeten we voor elke reële y nagaan welke gebeurtenis $A(y)$ voor $\underline{x}$ correspondeert met de gebeurtenis $(-\infty < \underline{y} \leq y)$ voor $\underline{y}$.
Dan geldt nl.:

$$G(y) := P[-\infty < \underline{y} \leq y] = P[\underline{x} \in A\ (y)].$$

Voorbeeld: $\underline{x}$ is een continue stochastische variabele met de reële getallenrechte tot uitkomstenruimte en distributiefunctie $F(x)$.
Beschouw nu de stochastische variabele $\underline{y} = \underline{x}^2$ met dezelfde uitkomstenruimte en met distributiefunctie $G(y)$.

Als $y < 0$ dan correspondeert met de gebeurtenis $[-\infty < \underline{y} \leq y]$
voor $\underline{y}$ de lege gebeurtenis voor $\underline{x}$;
Als $y \geq 0$ dan correspondeert met de gebeurtenis $[-\infty < \underline{y} \leq y]$
voor $\underline{y}$ de gebeurtenis $[-\sqrt{y} \leq \underline{x} \leq \sqrt{y}]$ voor $\underline{x}$.
Hieruit volgt:

$$G(y) = 0 \qquad \text{als } y < 0$$
$$G(y) = P[-\sqrt{y} \leq \underline{x} \leq \sqrt{y}]$$
$$= P[-\sqrt{y} < \underline{x} \leq \sqrt{y}]$$

(wegens de continuiteit van $F(x)$ is $P[\underline{x} = -\sqrt{y}] = 0$!)

$$= F(\sqrt{y}) - F(-\sqrt{y}) \qquad \text{als } y \geq 0.$$

Is $g(y)$ de dichtheidsfunctie van $\underline{y}$ en $f(x)$ die van $\underline{x}$, dan geldt;

$$g(y) = 0 \qquad \text{als } y < 0$$
$$g(y) = \frac{1}{2\sqrt{y}}\left(f(\sqrt{y}) + f(-\sqrt{y})\right) \qquad \text{als } y \geq 0$$

Bewijs!

Het komt vaak voor dat de functie $y = f(x)$ monotoon (niet dalend of niet stijgend) en differentieerbaar is. Dan is ook de inverse functie $\Psi = y(y)$ monotoon (niet dalend resp. niet stijgend) en differentieerbaar.
Zij nu $\underline{x}$ een stochastische grootheid met distributiefunctie $F(x)$ en dichtheidsfunctie $f(x)$.
Is $G(y)$ de distributiefunctie en $g(y)$ de dichtheidsfunctie van $\underline{y}$ dan geldt:

$$G(y) = P[\underline{y} \leq y] = P[\underline{x} \leq \Psi(y)] = F(\Psi(y)),$$
$$G(y) = \frac{d}{dy} G(y) = f(\Psi(y)) \cdot \left| \frac{d}{dy} \Psi(y) \right|$$

8. Twee dimensionale stochastieken

Tot nu toe hebben we het begrip distributie functie uitsluitend gebruikt voor stochastische variabelen $\underline{x}$ met een een-dimensionale uitkomstenruimte bestaande dus uit de verzameling van reele getallen of uit een eindige of aftelbaar oneindige deelverzameling daaruit. Dit begrip distributie functie kan ook bij stochastische variabelen met een meer dimensionale uitkomstenruimte gebruikt worden.
Zij $\underline{z} = (\underline{x},\underline{y})$ een stochastische variabele met het tweedimensionale (x,y) vlak tot uitkomstenruimte. Men kan nu weer aantonen dat de kansverdeling van $\underline{z} = (\underline{x},\underline{y})$ volledig bepaald is als de distributie functie

$$F(x,y) = P[\underline{x} \leq x,\ \underline{y} \leq y] ,$$

gedefinieerd voor elke $(x,y) \in U$, bekend is.

Is $\underline{z} = (\underline{x},\underline{y})$ een discrete stochastische variabele, is dus U een discrete verzameling $U = \{(x_i,y_i),\ i = 1,2,\ldots\}$ dan wordt

$$f(x,y) = P[\underline{x} = x_i,\ \underline{y} = y_i]$$

de frequentie functie van $(\underline{x},\underline{y})$ genoemd.

Is $\underline{z} = (\underline{x},\underline{y})$ een continue stochastische variabele met een twee maal differentieerbare distributie functie dan wordt

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

de kansdichtheid van $(\underline{x},\underline{y})$ genoemd.

De plausibiliteit van deze laatste naamgeving volgt weer uit

$$f(x,y) = \lim_{\substack{\Delta x \to o \\ \Delta y \to o}} \frac{F(x+\Delta x;\, y+\Delta y) - F(x,y)}{\Delta x \,.\, \Delta y}$$

$$= \lim_{\substack{\Delta x \to o \\ \Delta y \to o}} \frac{P[x < \underline{x} \leq x+\Delta x,\; y < \underline{y} \leq y+\Delta y]}{\Delta x \,.\, \Delta y}$$

Achter het limiet teken staat weer de kans op de gebeurtenis gerepresenteerd door het rechthoekje $(x,\ x+\Delta x;\ y,\ y+\Delta y)$ gelijkmatig uitgesmeerd over deze rechthoek.
Omgekeerd geldt in dit continue geval

$$F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x,y)\ dxdy$$

voor elke $(x,y) \in U$.

Uit de definitie van distributie functie, frequentie functie en dichtheidsfunctie, en uit de vermenigvuldigingsregel van kansen op gebeurtenissen betreffende onafhankelijke stochastische variabelen volgt:

Als $\underline{z} = (\underline{x},\underline{y})$ een stochastische variabele is met de tweedimensionale euclidische ruimte tot uitkomstenruimte terwijl de componenten $\underline{x}$ en $\underline{y}$ stochastisch onafhankelijk van elkaar zijn, dan geldt:

$$F(x,y) = F(x).F(y),$$

$$f(x,y) = f(x).f(y).$$

9.1 De verwachtingswaarde van een stochastische variabele

In voorgaande hoofdstukken hebben we gezien dat het stochastische gedrag van een variabele $\underline{x}$ volledig bekend is als zijn distributie functie bekend is.

Kennis van de kans op het voorkomen van elke mogelijke gebeurtenis is in vele practische situaties echter geen vereiste. Het blijkt dat bij herhaalde realisering van $\underline{x}$ de gevonden waarden x zich ophopen rond een vast punt van de uitkomstenruimte. Dit punt wordt de"verwachtingswaarde van $\underline{x}$" genoemd en kennis daarvan is bij vele toepassingen van primair belang. Het is echter ook van belang te weten binnen welke afstand tot deze verwachtingswaarde de gerealiseerde x-waarden met, zeg, kans 0.99 liggen. Is deze afstand klein dan zijn er slechts weinig realisaties nodig om de verwachtingswaarde met enige zekerheid te kunnen localiseren; is deze afstand groot dan kost dit localiseren vaak zeer veel realisaties.
Wij zullen dit alles nu nader wiskundig specificeren.

Zij $\underline{x}$ een discrete stochastische variabele met uitkomstenruimte $U = \{x_i,\ i = 1,2,\ldots\}$ en frequentie functie $f(x_i) = P(\underline{x} = x_i)$, $i = 1,2,\ldots$.
Onder de verwachtingswaarde $E(\underline{x})$ van $\underline{x}$ verstaat men dan:

$$E[\underline{x}] = \Sigma_i\ x_i\ f(x_i),$$

mits de som in het rechterlid bestaat.

Zij $\underline{x}$ een continue stochastische variabele met uitkomstenruimte de reele getallenrechte en met dichtheidsfunctie $f(x)$.
Onder de verwachtingswaarde $E(\underline{x})$ van $\underline{x}$ verstaat men dan:

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

mits deze integraal bestaat.

Zij $\varphi(x)$ een functie gedefinieerd op de uitkomstenruimte U van de stochastiche variabele $\underline{x}$. Onder de verwachtingswaarde van $\varphi(\underline{x})$ verstaat men dan (zelfde condities voor $\underline{x}$ als boven) in het discrete geval:

$$E[\varphi(\underline{x})] = \Sigma_i \varphi(x_i) f(x_i)$$

mits de reeks in het rechterlid absoluut convergeert;
in het continue geval:

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

mits de overeenkomstige integraal voor $|\varphi(x)|$ bestaat.

De verwachtingswaarde van $E[\varphi(\underline{x})]$ van een stochastische variabele $\varphi(\underline{x})$ heeft enkele zeer belangrijke eigenschappen welke bij discrete verdelingen terug te voeren zijn op eigenschappen van reeksen en bij continue verdelingen op elementaire eigenschappen van bepaalde integralen. Aangezien de bepaalde integralen in voorafgaande colleges slechts zeer summier behandeld zijn zullen wij ons bij de bewijsvoering tot het discrete geval beperken. De te noemen eigenschappen van $E(\underline{x})$ gelden echter algemeen; de enige voorwaarde is, dat de verwachtingswaarde van $|\varphi(\underline{x})|$ bestaat.

9.2 Eigenschappen van de verwachtingswaarde

Zij $\underline{x} = (\underline{y},\underline{z})$ een stochastische variabele met uitkomstenruimte U.
Zij V de uitkomstenruimte van $\underline{y}$ en W de uitkomstenruimte van $\underline{z}$.
Zij $\varphi(y)$ een functie gedefinieerd op V. De functie $\varphi(y)$ is dan ook op te vatten als een functie $\psi(x) = \psi(y,z) = \varphi(y)$ van $x = (y,z)$ gedefinieerd op W.
We kunnen dus de verwachtingswaarde van $\varphi(y) = \psi(y,z)$ berekenen uitgaande van de stochastische variabele $\underline{x} = (\underline{y},\underline{z})$ en ook uitgaande van de stochastische variabele $\underline{y}$. In beide gevallen vindt men echter dezelfde uitkomst:

$\underline{E}_1$ $\qquad E[\psi(y,z)] = E[\varphi(y)]$

want $\qquad E[\psi(y,z)] = \Sigma_{i,j} \psi(y_i,z_j) f(y_i,z_j)$

$$= \Sigma_{i,j} \varphi(y_i) f(y_i,z_j)$$

Wegens de absolute convergentie van de reeks in het rechterlid geldt

$$= \Sigma_i \varphi(y_i) \cdot \Sigma f(y_i, z_j)$$

$$= \Sigma_i \varphi(y_i) \cdot P[\underline{y} = y_i, \underline{z} \in W]$$

$$= \Sigma_i \varphi(y_i) \cdot P[\underline{y} = y_i]$$

$$= E[\varphi(\underline{y})] \cdot$$

Een practische interpretatie van deze stelling is: Als een verschijnsel gekarakteriseerd wordt door een stochastische variabele met twee of meer componenten en beschouwen wij een functie waarbij slechts één van de componenten als argument optreedt dan kunnen wij bij het bepalen van de verwachtingswaarde van die functie het stochastische gedrag van de overige componenten buiten beschouwing laten. We hoeven zelfs niet te weten dat zulke componenten aanwezig zijn!

Met de notatie $E[\varphi(\underline{x})]$ voegen we aan de stochastische variabele $\varphi(\underline{x})$ bij gegeven stochastisch gedrag van $\underline{x}$ de verwachtingswaarde toe. Men noemt E dan ook een operator.

De operator is lineair d.w.z. :

$\underline{E}_2$ Voor elk reëel getal λ geldt:

$$E[\lambda\varphi(\underline{x})] = \lambda\, E[\varphi(\underline{x})] \cdot$$

en:

$\underline{E}_3$ Als $\varphi(x)$ en $\psi(x)$ twee functies zijn, gedefinieerd op de uitkomstenruimte U van $\underline{x}$ dan is

$$E[\varphi(\underline{x}) + \psi(\underline{x})] = E[\varphi(\underline{x})] + E[\psi(\underline{x})].$$

Bewijs zelf voor het discrete geval.

$\underline{E}_4$ Zij $\underline{x} = (\underline{y},\underline{z})$ een stochastische variabele met uitkomsten ruimte U terwijl $\underline{y}$ de uitkomsten ruimte V en $\underline{z}$ de uitkomsten ruimte W heeft. Zij $\varphi(y)$ een functie gedefinieerd op V en $\psi(z)$ een functie gedefinieerd op W. Dan is $\varphi(y) + \psi(z)$ een functie gedefinieerd op U waarvoor geldt:

$$E[\varphi(\underline{y}) + \psi(\underline{z})] = E[\varphi(\underline{y})] + E[\psi(\underline{z})] .$$

Geven we door middel van de uitkomsten ruimte aan t.o.v. welke verdeling we de verwachtingswaarde bepalen dan geldt:

$$E_u[\varphi(\underline{y}) + \psi(\underline{z})] = E_u[\varphi(\underline{y})] + E_u[\psi(\underline{z})] \quad \text{wegens eigenschap } \underline{E}_3$$

$$= E_v[\varphi(\underline{y})] + E_w[\psi(\underline{z})] \quad \text{wegens eigenschap } \underline{E}_1$$

Men merke op dat voor de geldigheid van eigenschap $\underline{E}_1$ en dus van eigenschap $\underline{E}_4$ niet vereist is dat de stochastische variabelen $\underline{y}$ en $\underline{z}$ onafhankelijk zijn.
Dit is wel het geval bij de volgende stelling.

$\underline{E}_5$ Als $\underline{y}$ en $\underline{z}$ twee onafhankelijke stochastische variabelen zijn terwijl $\varphi(y)$ een functie is gedefinieerd op de uitkomsten ruimte V van $\underline{y}$ en $\psi(z)$ op de uitkomsten ruimte W van $\underline{z}$ dan geldt:

$$E[\varphi(\underline{y}).\psi(\underline{z})] = E[\varphi(\underline{y})].\ E[\psi(\underline{z})].$$

Want

$$E[\varphi(\underline{y}).\psi(\underline{z})] = \Sigma_{ij}\ \varphi(y_i)\ \psi(z_j)\ f(y_i,z_j)$$

$$= \Sigma_{ij}\ \varphi(y_i)\ \psi(z_j)\ f(y_i)\ f(z_j)$$

(onafhankelijkheid van $\underline{y}$ en $\underline{z}$)

$$= \Sigma_{ij}\ \varphi(y_i)\ f(y_i)\ \psi(z_j)\ f(z_j)$$

$$= \Sigma_i\ \varphi(y_i)\ f(y_i)\ \ \Sigma_j\ \psi(z_j)\ f(z_j)$$

(absolute convergentie van reeksen)

$$= E\ \varphi(\underline{y})\ E\psi(\underline{z}).$$

Verwachtingswaarde van enkele vaak voorkomende stochastische variabelen

a. Bernoulli stochastiek

stochastische variabele $\underline{x}$: uitkomstengebied $U = (0,1)$

$P(\underline{x}=1) = p$, $P(\underline{x}=0) = 1-p$

$$E(\underline{x}) = p.1 + (1-p).0 = p.$$

b. Binomiale stochastiek

stochastische variabele $\underline{y} = \sum_{i=1}^{n} \underline{x}_i$; $\underline{x}_i$ bernoulli stochastiek met $P(\underline{x}_i = 1) = p$ en $P(\underline{x}_i = 0) = 1-p$ voor alle $i = 1,\ldots,n$; uitkomstengebied van $\underline{y}$ is $U = (0,1,\ldots,n)$; $P(\underline{y}=m) = \binom{n}{m} p^m(1-p)^{n-m}$, $m = 0,\ldots,n$.

$$E(\underline{y}) = \sum_{m=0}^{n} m.\binom{n}{m}p^m(1-p)^{n-m} = np.$$

Leidt dit resultaat ook af met behulp van eigenschap E_3, pag 38, uit de verwachtingswaarde van de bernoulli stochastiek.

c. Poisson stochastiek

stochastische variabele $\underline{x}$; uitkomstengebied $U = (0,1,2,\ldots)$

$f(x) = P(\underline{x} = x) = e^{-\lambda}\frac{\lambda^x}{x!}$, $x = 0,1,\ldots$.

$$E(\underline{x}) = \sum_{x=0}^{\infty} x.e^{-\lambda}\frac{\lambda^x}{x!} = \lambda .$$

De parameter λ, voorkomend in de frequentie functie van de poisson stochastiek $\underline{x}$ is dus gelijk aan de verwachtingswaarde van $\underline{x}$.

d. De continue uniforme stochastiek

stochastische variabele $\underline{x}$; uitkomstengebied het interval $U(a,b)$

$f(x) = \frac{1}{b-a}$ voor $a \leqslant x \leqslant b$

$$E(\underline{x}) = \int_a^b \frac{x}{b-a}\,dx = \tfrac{1}{2}(a+b)$$

e. De negatief exponentiële stochastiek

stochastische variabele $\underline{x}$; uitkomstenruimte $U = (x \geq 0)$

$f(x) = \lambda e^{-\lambda x}$, $x \geq 0$

$$E(\underline{x}) = \int_0^{\infty} x.\lambda e^{-\lambda x} dx = \frac{1}{\lambda} .$$

f. Normale stochastiek

stochastische variabele $\underline{x}$; uitkomstenruimte $U = (-\infty < x < +\infty)$,

$$P(\underline{x} = x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

$$E(\underline{x}) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} x\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \mu .$$

Bij een normale verdeling is de parameter μ, optredende in de frequentieverdeling van $\underline{x}$ gelijk aan de verwachtingswaarde van $\underline{x}$.

De verwachtingswaarde is een z.g. "centrale parameter" van een distributiefunctie: hij localiseert het meest belangrijke gebied van de uitkomstenruimte. Deze uitspraak zal in hoofdstuk 11 nog nader worden gepreciseerd.

10.1 De spreiding en de variantie van een stochastische variabele

Naast een centrale parameter voor het localiseren van het meest belangrijkste gebied van de uitkomstenruimte wensen we ook een maat te hebben voor de uitgestrektheid van dit gebied.
Zo'n gebied zou men als volgt kunnen afbakenen: Men kiest een positief getal, bv. $\alpha = 0.05$ of $\alpha = 0.01$. Daarna bepaalt men het grootste getal r waarvoor

$$P\left[E(\underline{x}) - r \leq \underline{x} \leq E(\underline{x}) + r\right] \leq 1 - \alpha .$$

De rol die $E(\underline{x})$ hierbij als centrale parameter speelt is duidelijk. Het zal in het volgende blijken dat de grootheid

$$\sigma(\underline{x}) = \sqrt{E(\underline{x} - E(\underline{x}))^2}$$

een betere hanteerbare maat is voor de uitgestrektheid van het meest belangrijke gebied van de uitkomstenruimte.

$\sigma(\underline{x})$ heet de "spreiding" van $\underline{x}$. (Alternatieve termen zijn "standaarddeviatie", "standaardfout").
$\sigma^2(\underline{x}) = E(\underline{x} - E(\underline{x}))^2$ heet de variantie van $\underline{x}$.

Er zijn stochastieken $\underline{x}$ waarvoor $\sigma(\underline{x})$ niet bestaat. Tenzij anders vermeld zullen wij echter alleen stochastieken $\underline{x}$ beschouwen waarvoor $\sigma(\underline{x})$ wel bestaat.

10.2 Eigenschappen van de variantie

$\underline{V}_1$ De variantie $\sigma^2(\underline{x})$ van een stochastische variabele $\underline{x}$ is dan en slechts dan gelijk 0 als er een punt x_o is in de uitkomstenruimte U waarvoor geldt $P[\underline{x} = x_o] = 1$.
Men zegt dan dat de totale waarschijnlijkheid in het punt x_o geconcentreerd is.
Bewijs zelf.

$\underline{V}_2$ $\sigma^2(\underline{x}+a) = \sigma^2(\underline{x})$ d.w.z. de stochastieken $\underline{x}$ en $\underline{y} = \underline{x} + a$ hebben dezelfde variantie.
Uit $E(\underline{x} + a) = E(\underline{x}) + a$ volgt

$$\sigma^2(\underline{x} + a) = E(\underline{x}+a - E(\underline{x} + a))^2$$
$$= E(\underline{x}+a - E(\underline{x}) - a)^2 = E(\underline{x} - E(\underline{x}))^2 = \sigma^2(\underline{x}).$$

$\underline{V}_3$ $\sigma^2(a\underline{x}) = a^2\sigma^2(\underline{x})$ d.w.z. vermenigvuldigt men een stochastische variabele $\underline{x}$ met a dan wordt zijn variantie met a^2 vermenigvuldigd.
Bewijs zelf. Formuleer deze eigenschap ook in termen van standaard afwijkingen.

$\underline{V}_4$ Als $\underline{x} = (\underline{y},\underline{z})$ een stochastische variabele is met de twee dimensionale euclidische ruimte tot uitkomstenruimte, dan geldt

$$\sigma^2(\underline{y} + \underline{z}) = \sigma^2(\underline{y}) + \sigma^2(\underline{z})$$

onder voorwaarde dat $\underline{y}$ en $\underline{z}$ onafhankelijk van elkaar zijn.
Dit blijkt als volgt:

$$\begin{aligned}\sigma^2(\underline{y}+\underline{z}) &= E[\underline{y}+\underline{z}-E(\underline{y}+\underline{z})]^2\\ &= E[(\underline{y}-E(\underline{y}))+\underline{z}-E(\underline{z})]^2\\ &= E(\underline{y}-E(\underline{y}))^2 + E(\underline{z}-E(\underline{z}))^2 + 2\,E(\underline{y}-E(\underline{y}))(\underline{z}-E(\underline{z}))\\ &= E(\underline{y}-E(\underline{y}))^2 + E(\underline{z}-E(\underline{z}))^2\\ &= \sigma^2(\underline{y}) + \sigma^2(\underline{z}).\end{aligned}$$

Algemener geldt

$$\sigma^2(\underline{y}+\underline{z}) = \sigma^2(\underline{y}) + \sigma^2(\underline{z}) + 2\,\mathrm{Cov}\,(\underline{y},\underline{z}).$$

Onder de covariantie van twee stochastische grootheden $\underline{y}$ en $\underline{z}$ verstaat men

$$\mathrm{Cov}\,(\underline{y},\underline{z}) = E[\underline{y}-E(\underline{y})][\underline{z}-E(\underline{z})]\ .$$

$\underline{V}_5$ De covariantie van twee onafhankelijke stochastische grootheden $\underline{y}$ en $\underline{z}$ is gelijk nul.
Volgt uit $\underline{E}_5$ en $E(\underline{y}-E(\underline{y})) = E(\underline{y})-E(\underline{y}) = 0$.
Het omgekeerde van $\underline{V}_5$ geldt <u>niet</u>: er zijn voorbeelden te construeren van afhankelijke stochastische variabelen $\underline{x}$ en $\underline{y}$ met covariantie gelijk aan 0. (Weten we echter dat de stochastieken $\underline{x}$ en $\underline{y}$ normaal zijn, dan geldt dit omgekeerde wel. We bewijzen deze laatste uitspraak hier echter niet.)

$\underline{V}_6$ Als $\{\underline{x}_i,\ i=1,\ldots,n\}$ een n-tal onderling onafhankelijke stochastische variabelen is dan geldt

$$E[\Sigma_i\, x_i] = n\,E(\underline{x}_i) = n\,E(\underline{x})$$
$$\sigma^2[\Sigma_i\, \underline{x}_i] = \Sigma_i\, \sigma^2[\underline{x}_i] = n\,\sigma^2(\underline{x})$$
en $$\sigma[\Sigma_i\, x_i] = \sigma(\underline{x})\sqrt{n}.$$

In woorden: De stochastische variabele $\underline{y}$ welke ontstaat door het optellen van n onderling onafhankelijke maar gelijk verdeelde stochastische variabelen $\underline{x}_i$, $i = 1,\ldots,n$, (verwachtingswaarde $E(\underline{x})$, spreiding $\sigma(\underline{x})$ en variantie $\sigma^2(\underline{x})$) heeft een verwachtingswaarde welke gelijk is aan n keer die van $\underline{x}_i$, een variantie welke gelijk is aan n keer die van $\underline{x}_i$, en een spreiding welke gelijk is aan $\sqrt{n}$ keer die van $\underline{x}_i$.

Varianties van enkele stochastische variabelen

a. Bernoulli stochastiek

stochastische variabele $\underline{x}$; uitkomstenruimte $U = (0,1)$

$P(\underline{x} = 1) = p$; $P(\underline{x} = 0) = 1 - p$;

Verwachtingswaarde $E(\underline{x}) = p$

Variantie $\sigma^2(\underline{x}) = E(\underline{x} - p)^2 = p(1 - p)$

b. Binomiale stochastiek

stochastische variabele $\underline{y} = \sum_{i=1}^{n} \underline{x}_i$; uitkomstenruimte $U = (0,1,\ldots,n)$

$\underline{x}_i$ bernoulli stochastiek met $P(\underline{x} = 1) = p$ en $P(\underline{x} = 0) = 1 - p$.

Verwachtingswaarde $E(\underline{y}) = np$

Variantie $\sigma^2(\underline{y}) = E(\underline{y} - np)^2 = np(1 - p)$.

Leid dit resultaat ook af met behulp van eigenschap V_4, pag. 31, uit de variantie van de bernoulli stochastiek.

c. Poisson stochastiek

stochastische variabele $\underline{x}$; uitkomstenruimte $U = (0,1,2,\ldots,)$

Frequentieverdeling $f(x) = e^{-\lambda} \frac{\lambda^x}{x!}$, $x = 0,1,\ldots$.

Verwachtingswaarde $E(\underline{x}) = \lambda$

Variantie $\sigma^2(\underline{x}) = E(\underline{x} - \lambda)^2 = \lambda$.

Bij een poisson verdeling met parameter λ, is deze parameter gelijk aan de verwachtingswaarde en aan de variantie. De poisson verdeling is dus zowel door zijn verwachtingswaarde als door zijn variantie volledig bepaald.

d. De continue uniforme stochastiek

Stochastische variabele $\underline{x}$; uitkomstenruimte het interval $U(a \leq x \leq b)$

Dichtheidsfunctie $f(x) = \frac{1}{b - a}$ voor $a \leq x \leq b$.

Verwachtingswaarde $E(\underline{x}) = \frac{1}{2}(a + b)$

Variantie $\sigma^2(\underline{x}) = E(\underline{x} - \frac{1}{2}(a + b))^2 = \frac{1}{12}(b - a)^2$

e. Exponentiële stochastiek

stochastische variabele $\underline{x}$; uitkomstenruimte $U = (x \geq 0)$

Dichtheidsfunctie $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$

Verwachtingswaarde $E(\underline{x}) = \frac{1}{\lambda}$

Variantie $\sigma^2(\underline{x}) = E(\underline{x} - \frac{1}{\lambda})^2 = \frac{1}{\lambda^2}$.

f. Normale stochastiek

stochastische variabele $\underline{x}$; uitkomstenruimte $U = (-\infty < x < \infty)$

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

Verwachtingswaarde $E(\underline{x}) = \mu$

Variantie $\sigma^2(\underline{x}) = E(\underline{x} - \mu)^2 = \sigma^2$

De normale verdeling is door verwachtingswaarde en variantie volledig bepaald.

De twee-sigma en drie-sigma regels voor de normale verdeling

De normale verdeling is op vele plaatsen getabelleerd voor het speciale geval $\mu = 0$, $\sigma = 1$. (gestandaardiseerde normale verdeling)
(Ga na hoe hieruit een tabel voor de normale verdeling voor willekeurige μ en σ kan worden afgeleid.)
Uit deze tabellen blijkt:

$$P(\mu - 2\sigma \leq \underline{x} \leq \mu + 2\sigma) = 0{,}9544$$

$$P(\mu - 3\sigma \leq \underline{x} \leq \mu + 3\sigma) = 0{,}9974$$

waaruit men voor practische toepassingen de volgende vuistregels kan afleiden:

De twee-sigma regel: De kans dat bij een normale verdeling een realisatie x verder dan 2σ van de verwachtingswaarde verwijderd is, is 0.05.

De drie-sigma regel: de kans dat bij een normale verdeling een realisatie x verder dan 3σ van de verwachtingswaarde verwijderd is, is 0.003.

11 Wetten van de grote aantallen

In het vorige hoofdstuk werden de verwachtingswaarde en de variantie van een stochastische variabele genoemd als voornaamste grootheden voor het min of meer nauwkeurig karakteriseren van het stochastisch gedrag van zo'n variabele. We gaan deze beweringen hier nauwkeurig specificeren.

We zullen ons beperken tot een-dimensionale stochastische variabelen $\underline{x}$.
De uitkomstenruimte is dus de verzameling van reële getallen of een deelverzameling daarvan. Verder nemen we aan dat de distributie functie $F(x) = P(\underline{x} \leq x)$ van $\underline{x}$ bekend is, in het discrete geval ook de frequentie verdeling $f(x) = P(\underline{x} = x)$ en in het continue geval de dichtheids functie $f(x) = \frac{d}{dx} F(x)$

WGA 1. De formule van Bienaymé - Cebysev

Als $\underline{x}$ een stochastische variabele is waarvoor zowel de verwachtingswaarde $E(\underline{x})$ als de variantie $\sigma^2(\underline{x})$ bestaan, dan geldt voor elke $a > 0$:

$$P[|\underline{x} - E(\underline{x})| \geq a] \leq \frac{\sigma^2(\underline{x})}{a^2}.$$

Dus, de kans op de gebeurtenis $|\underline{x} - E(\underline{x})| \geq a$, d.w.z. de kans dat bij realisering van de stochastische variabele $\underline{x}$ deze een waarde krijgt die minstens het bedrag a van de verwachtingswaarde $E(\underline{x})$ verwijderd ligt, is beperkt en niet groter dan $\frac{\sigma^2(\underline{x})}{a^2}$. Uit het feit dat a in de noemer van deze bovengrens staat volgt dat deze kans kleiner wordt naarmate a groter wordt. Uit deze formule blijkt zowel de betekenis van $E(\underline{x})$ als centrale parameter en van de spreiding $\sigma(\underline{x})$ als maat voor de uitgestrektheid van het meest interessante deel van de uitkomstenruimte. Onderstaande tabel geeft de kans op de gebeurtenis $[|\underline{x} - E(\underline{x})| \geq a]$, als voor a veelvouden van $\sigma(\underline{x})$ genomen worden.

Het gebruik van de formule van Bienaymé - Cebysev veronderstelt geen informatie over de verdelingsfunctie van $\underline{x}$. Is zulke informatie wel aanwezig dan kunnen veel scherpere uitspraken gedaan worden. Ter illustratie daarvan zijn in de derde kolom van onderstaande tabel de kansen op afwijkingen van $E(\underline{x})$ groter dan a opgenomen als bekend is dat $\underline{x}$ een normale stochastiek is.

a	Bienaymé - Cebysev	Normale stochastiek
$\sigma(\underline{x})$	$P(\lvert\underline{x} - E(\underline{x})\rvert \geq a) \leq 1$	$P(\lvert\underline{x} - E(\underline{x})\rvert \geq a) \leq 0.32$
$2\ \sigma(\underline{x})$	≤ 0.25	≤ 0.05
$3\ \sigma(\underline{x})$	≤ 0.11	≤ 0.003
$4\ \sigma(\underline{x})$	≤ 0.06	≤ 0.000
$5\ \sigma(\underline{x})$	≤ 0.04	≤ 0.000
$10\ \sigma(\underline{x})$	≤ 0.01	≤ 0.000

We zullen het bewijs geven van de formule van Bienaymé - Cebysev als $\underline{x}$ een continue stochastische variabele is.

$$\sigma^2(\underline{x}) = \int_{-\infty}^{\infty} (\underline{x} - E(\underline{x}))^2 f(x)\,dx$$

$$= \int_{-a+E(\underline{x})}^{a+E(\underline{x})} (\underline{x} - E(\underline{x}))^2 f(x)\,dx +$$

$$+ \int_{-\infty}^{-a+E(\underline{x})} (\underline{x} - E(\underline{x}))^2 f(x)\,dx + \int_{a+E(\underline{x})}^{\infty} (\underline{x} - E(\underline{x}))^2 f(x)\,dx$$

$$\geqslant a^2 \{ \int_{-\infty}^{-a+E(\underline{x})} f(x)dx + \int_{a+E(\underline{x})}^{\infty} f(x)dx \}$$

$$\geqslant a^2\, P(|\underline{x} - E(\underline{x})| \geqslant a)$$

waaruit volgt:

$$P(|\underline{x} - E(\underline{x})| \geqslant a) \leqslant \frac{\sigma^2(\underline{x})}{a^2}.$$

(Geef zelf het bewijs in het geval $\underline{x}$ een discrete stochastiek is.)

De samenhang tussen het rekenkundig gemiddelde en de verwachtingswaarde van een stochastiek.

Laat x_i, $i = 1,\ldots,n$ een stel van n onafhankelijke realisaties voorstellen van een stochastische variabele $\underline{x}$ met verwachtingswaarde $E(\underline{x})$ en variantie $\sigma^2(\underline{x})$.
In practische gevallen is het gebruikelijk het rekenkundig gemiddelde $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ als "beste schatting van het bestudeerde verschijnsel te nemen" terwijl we in voorgaande hoofdstuk gesuggereerd hebben dat de verwachtingswaarde $E(\underline{x})$ zo'n "beste schatting" is. Met de formule van Bienaymé - Cebysev blijkt dat beide uitspraken gerechtvaardigd zijn. Nu geldt nl.

$$P[|\frac{1}{n} \Sigma\, \underline{x}_i - E(\underline{x})| \geqslant a] \leqslant \frac{1}{n} \cdot \frac{\sigma^2(\underline{x})}{a^2},$$

d.w.z. naarmate n groter wordt, is de kans kleiner (evenredig met $\frac{1}{n}$) dat het rekenkundig gemiddelde van onafhankelijke realisaties van $\underline{x}$ meer dan het bedrag a van de verwachtingswaarde $E(\underline{x})$ afwijkt.

(Bewijs deze formule zelf! Bedenk dat $\bar{x}$ een realisatie is van de nieuwe stochastische variabele $\underline{\bar{x}} = \frac{1}{n} \underline{x}_i$; ga na wat zijn variantie is; realiseer U waar de onafhankelijkheid van de realisaties x_i, $i = 1,\ldots,n$ gebruikt wordt.)

Gestandaardiseerde stochastieken

Zij $\underline{x}$ een stochastiek met verwachtingswaarde $E(\underline{x})$ en spreiding $\sigma(\underline{x})$.
De stochastiek $\underline{y}$ welke uit $\underline{x}$ kan worden afgeleid volgens

$$\underline{y} = \frac{\underline{x} - E(\underline{x})}{\sigma(\underline{x})}$$

heeft een verwachtingswaarde gelijk aan nul en een spreiding gelijk aan 1.
Deze stochastiek wordt de gestandaardiseerde vorm van de stochastiek $\underline{x}$ genoemd.

Voorbeelden

Binomiale stochastiek $\underline{x} = \sum_{i=1}^{n} \underline{x}_i$; $P[\underline{x}_i = 1] = p$
$P[\underline{x}_i = 0] = 1 - p$.

Gestandaardiseerde binomiale stochastiek:

$$\underline{z} = \frac{\sum_{i=1}^{n} \underline{x}_i - np}{\sqrt{np(1-p)}}$$

Poisson stochastiek $\underline{x}$; verwachtingswaarde $E(\underline{x}) = \lambda$.
Gestandaardiseerde poisson stochastiek:

$$\underline{y} = \frac{\underline{x} - \lambda}{\sqrt{\lambda}} \quad .$$

Normale stochastiek $\underline{x}$; verwachtingswaarde $E(\underline{x}) = \mu$;
spreiding $\sigma(\underline{x}) = \sigma$.

Gestandaardiseerde normale stochastiek:

$$\underline{y} = \frac{\underline{x} - \mu}{\sigma}.$$

De distributie functie van de gestandaardiseerde normale stochastiek wordt gewoonlijk aangeduid met $\Phi(x)$:

$$P[\underline{x} \leq x] = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt.$$

In plaats van de uitdrukking: $\underline{x}$ is een gestandaardiseerde normale stochastiek gebruikt men ook de uitdrukking: $\underline{x}$ is normaal verdeeld $(0,1)$ of nog korter: $\underline{x}$ is $N(0,1)$ verdeeld.
Hierin betekent 0 dan $E(\underline{x}) = 0$ en 1 dan $\sigma(\underline{x}) = 1$
Algemener $\underline{x}$ is $N(\mu, \sigma)$ verdeeld betekent dat $\underline{x}$ een normale stochastiek is met verwachtingswaarde μ en spreiding σ.
De distributie functie $\Phi(x)$, ook de standaard normale distributie functie genoemd, is in vele statistische handboeken en tabellenboeken getabelleerd.

De volgende wetten van de grote aantallen WGA 2, WGA 4 en WGA 5 karakteriseren de centrale positie welke de normale stochastiek inneemt in statistische beschouwingen.

WGA 2 De distributie functie van de gestandaardiseerde binomiale stochastiek convergeert voor vaste p en $n \to \infty$ naar de standaard normale distributie functie:
Als $\underline{x}_i$, $i = 1,\ldots,n$ een stel onafhankelijke bernoulli stochastieken is met $P(\underline{x}_i = 1) = p$, $P(\underline{x}_i = 0) = 1 - p$, dan geldt

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^{n} \underline{x}_i - np}{\sqrt{np(1-p)}} \leq x \right) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt.$$

De benadering van de distributie functie van de gestandaardiseerde binomiale stochastiek door de standaard normale distributie functie blijkt voor practische doeleinden bevredigend als

$$np(1 - p) > 9$$

is.

Onderstaande tabel geeft voor verschillende waarden van p de bijbehorende ondergrens voor n (afgerond op vijfvouden)

p	n	
0.5	35	
0.4	40	$np(1-p) \sim 9$
0.3	45	
0.2	55	
0.15	70	
0.10	100	
0.05	190	
0.02	450	
0.01	900	

Is in een practische situatie aan deze voorwaarde niet voldaan dan is men bij gebruik van de binomiale stochastiek aangewezen op tabellen voor de binomiale distributie functie. Deze zijn in de meeste statistische hand- en tabellenboeken te vinden. Wordt n groot en p klein (vuistregel $p \leq 0.1$) dan is het meestal veel handiger om gebruik te maken van de volgende limietstelling.

WGA 3. Als 1. $\underline{x}_i$, $i = 1,\ldots,n$ een stel onafhankelijke bernoulli stochastieken is met $P(\underline{x}_i = 1) = p$ en $P(\underline{x}_i = 0) = 1 - p$.

2. $\underline{y} = \sum_{i=1}^{n} \underline{x}_i$ de bijbehorende binomiale stochastiek is.

3. $E(\underline{y}) = np = \lambda$

dan geldt voor vaste λ en n groot

$$P(\underline{y} = y) = \binom{n}{y} p^y (1-p)^{n-y} \sim e^{-\lambda} \frac{\lambda^y}{y!}.$$

De eis: λ vast en n groot betekent dat p klein moet zijn. Men kan WGA 3 dan ook als volgt interpreteren:
Als p klein is en n groot, terwijl $np = \lambda$ dan wordt de binomiale distributie functie benaderd door de distributie functie van de poisson stochastiek met parameter λ .

Uit tabellen van de binomiale verdeling en van de poisson verdeling blijkt dat als

$$p \leq 0.1,$$

de distributie functie van de niet-gestandaardiseerde binomiale stochastiek voor kleine waarden van n ($n \sim 10$) reeds goed door de poisson verdeling met parameter $\lambda = np$ benaderd wordt.

Voor grotere waarden van n en p is het niet belangrijk of men de binomiale verdeling door een poisson verdeling dan wel door een normale verdeling benadert. Dit blijkt uit de volgende eigenschap:

WGA 4. Is $\underline{x}$ een poisson stochastiek met verwachtingswaarde $E(\underline{x}) = \lambda$, dan convergeert de distributie functie van de gestandaardiseerde poisson stochastiek $(\underline{x} - \lambda)/\sqrt{\lambda}$ voor $\lambda \to \infty$ naar de standaard normale distributie functie:

$$\lim_{\lambda \to \infty} P\left[\frac{\underline{x} - \lambda}{\sqrt{\lambda}} \leq x\right] = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt.$$

Uit tabellen van de normale distributie en de poisson distributie blijkt dat, voor practische doeleinden, de poisson verdeling vervangen mag worden door de normale verdeling als $\lambda > 9$ is.

De wetten van de grote aantallen WGA 2 en WGA 3 zijn bijzondere gevallen van:

WGA 5. <u>De centrale limietstelling</u>:
Onder zeer ruime voorwaarden geldt:
Als $\underline{x}_i$, $i = 1,\ldots,n$ een stel onderling onafhankelijke stochastieken is met verwachtingswaarde $\mu_i = E(\underline{x}_i)$ en spreidingen $\sigma_i = \sigma(\underline{x}_i)$, dan convergeert de distributie functie van de stochastische variabele

$$\underline{y} = \frac{\sum_{i=1}^{n} \underline{x}_i - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}}$$

voor $n \to \infty$ naar de standaard normale distributie functie $\Phi(x)$:

$$\lim_{n\to\infty} P\left[\frac{\sum_{i=1}^{n} \underline{x}_i - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}} \leq x\right] = \varphi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt.$$

Deze centrale limietstelling legaliseert de vaak gemaakte veronderstelling dat de stochastiek $\underline{x}$, geassocieerd aan een meting, een normale stochastiek is. Het resultaat van een meting is doorgaans namelijk de resultante van een te meten effect waarop een groot aantal toevallige storingen zijn gesuperponeerd die alle onafhankelijk van elkaar werken (meetfout, instelfout, afwijking van gewenste temperatuur, onzuiverheid van cuvetten enz.) Men bedenke echter dat deze redenering niet in alle gevallen opgaat: er zijn ook metingen waarmee een niet normale stochastiek verbonden is. Het is een samenspel van ervaring en critische instelling dat de statisticus hier de juiste keuze doet maken.

12. Steekproeven

In voorgaande hoofdstukken is reeks enkele keren gesteld dat met elke meting een stochastische variabele geassocieerd kan worden. Men neemt hierbij aan dat de te meten grootheid een "werkelijke waarde" heeft die samenvalt met de verwachtingswaarde $E(\underline{x})$ van deze stochastische variabele. De meting van deze waarde wordt echter beinvloed door een aantal toevallige storingen zoals instelfouten van apparatuur, afwijkingen van gewenste temperatuur, onzuiverheden van gebruikte stoffen, enz. Deze storingen bepalen de spreiding $\sigma(\underline{x})$.

Informatie over de "werkelijke waarde" van de te meten grootheid kan slechts verkregen worden door meting, dus door het één of meerdere malen realiseren van de stochastische variabele $\underline{x}$. Zo'n realisering noemt men "het uitvoeren van een steekproef". n keer realiseren van $\underline{x}$ levert een steekproef ter grootte n en de gevonden uitkomsten x_i, $i = 1,\ldots,n$ heten de steekproefuitkomsten. We nemen nu aan dat de spreiding $\sigma(\underline{x})$ van de stochastiek $\underline{x}$ bekend is, en dat de realisaties x_i, $i = 1, \ldots, n$ van $\underline{x}$ onafhankelijk van elkaar plaats hebben.

Het rekenkundig gemiddelde $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ is een realisatie van de stochastiek $\underline{\bar{x}} = \frac{1}{n} \sum_{i=1}^{n} \underline{x}_i$ waarvoor volgens Bienaymé-Cebysev bij gegeven a geldt

$$P\left(|\underline{\bar{x}} - E(\underline{x})| \geq a \right) \leq \frac{\sigma^2(\underline{x})}{na^2} .$$

Aangezien deze kans omgekeerd evenredig is met de steekproefgrootte kan men de steekproef zo groot nemen dat de kans dat het steekproefgemiddelde (dus dat een realisatie $\bar{x}$ van $\underline{\bar{x}}$) meer dan een voorafgegeven bedrag a afwijkt van de verwachtingswaarde $E(\underline{x})$, zo klein is als men zelf wil.

Men noemt daarom de stochastische variabele $\underline{\bar{x}} = \frac{1}{n} \sum_{i=1}^{n} \underline{x}_i$ een schatter van $E(\underline{x})$ en een realisatie $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ van de schatter een schatting van $E(\underline{x})$.

De formule van Bienaymé-Cebysev kan ook als volgt geïnterpreteerd worden:
De kans dat de uitspraak: bij gegeven bedrag a en een steekproef ter grootte n geldt

$$\bar{x} - a \leq E(\underline{x}) \leq \bar{x} + a$$

niet juist is, is hoogstens gelijk aan $\frac{\sigma^2(\underline{x})}{na^2}$.

Men noemt het interval

$$(\bar{x} - a, \bar{x} + a)$$

een betrouwbaarheidsinterval voor $E(\underline{x})$ behorende bij een onbetrouwbaarheid $\frac{\sigma^2(\underline{x})}{na^2}$.
Deze onbetrouwbaarheid is kleiner dan een gegeven bedrag α als $\frac{\sigma^2(\underline{x})}{na^2} < \alpha$, dus als de steekproefgrootte $n > \frac{\sigma^2(\underline{x})}{\alpha a^2}$ is.
In plaats van "betrouwbaarheidsinterval" zegt men ook wel "confidentie interval".

De formule van Bienaymé-Cebysev geldt onafhankelijk van de distributie functie welke bij de stochastische $\underline{x}$ behoort. Dit heeft tot gevolg, dat de daarop gebaseerde confidentie intervallen groter zullen uitvallen dan wanneer meer informatie over deze distributie functie beschikbaar is. (Zie tabel pag. 46) We hebben reeds gezien dat met vele metingen een normale stochastiek $\underline{x}$ geassocieerd is. (Op grond van de centrale limiet stelling.) Laat x_i, i = 1, ..., n een

steekproef ter grootte n zijn van een normale stochastiek met verwachtingswaarde $E(\underline{x}) = \mu$ en spreiding $\sigma(\underline{x}) = \sigma$. (onderling onafhankelijke realisaties)
We weten dan dat het rekenkundig gemiddelde $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ een steekproef ter grootte 1 is van de stochastiek

$$\underline{\bar{x}} = \frac{1}{n} \sum_{i=1}^{n} \underline{x}_i$$

welke een verwachtingswaarde $E(\underline{y}) = E(\underline{x}) = \mu$ heeft en een spreiding

$$\sigma_{\underline{y}} = \frac{1}{\sqrt{n}} \sigma$$

heeft.
Men kan zelfs bewijzen dat $\underline{y}$ weer een normale stochastiek is.

Uit de twee-sigma regel (zie pag 34) volgt nu:
Bij een onbetrouwbaarheid van 0.05 geldt voor de verwachtingswaarde $\mu = E(\underline{x})$ het confidentie interval

$$\left(\frac{1}{n} \sum_{i=1}^{n} x_i - \frac{2}{\sqrt{n}} \sigma, \frac{1}{n} \sum_{i=1}^{n} x_i + \frac{2}{\sqrt{n}} \sigma\right);$$

Uit de drie-sigma regel (zie pag 34) volgt:
Bij een onbetrouwbaarheid van 0.003 geldt voor de verwachtingswaarde $\mu = E(\underline{x})$ het confidentie interval:

$$\left(\frac{1}{n} \sum_{i=1}^{n} x_i - \frac{3}{\sqrt{n}} \sigma, \frac{1}{n} \sum_{i=1}^{n} x_i + \frac{3}{\sqrt{n}} \sigma\right).$$

Universum grootheden en Steekproef grootheden

Bij statistische beschouwingen dient steeds een scherp onderscheid gemaakt te worden tussen universum grootheden, de "echte" grootheden waarover men informatie wil hebben, en steekproef grootheden op grond waarvan conclusies worden getrokken.

universum	steekproef
universum gemiddelde verwachtingswaarde $\mu = E(\underline{x})$	steekproefgrootte: n steekproefuitkomsten: $x_1,\ldots,x_n$. steekproefgemiddelde $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
universum variantie $\sigma^2 = \sigma^2(\underline{x})$	steekproefvariantie $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
universum spreiding $\sigma = \sigma(\underline{x})$	steekproefspreiding $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

De factor n - 1 in de noemer van de steekproefvariantie vereist enige toelichting. Men zou daar nl. een factor n verwacht hebben. Uit de definitie van $\sigma^2(\underline{x})$, pag.41 en de formule van Bienaymé - Cebysev volgt dat de grootheid

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - E(\underline{x}))^2$$

een schatting is van de universum variantie $\sigma^2(\underline{x})$.
Nu is de verwachtingswaarde $E(\underline{x})$ in het algemeen niet bekend. Vervangt men echter het universum gemiddelde $E(\underline{x})$ door het steekproefgemiddelde $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ en in de noemer n door n - 1, dan kan men aantonen dat ook

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

een schatting is van de variantie $\sigma^2(\underline{x})$.
Dit blijkt als volgt:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}((x_i - E(\underline{x}) - (\bar{x} - E(\underline{x}))^2$$

$$= \sum_{i=1}^{n} (x_i - E(\underline{x}))^2 + \sum_{i=1}^{n} (\bar{\underline{x}} - E(\underline{x}))^2 - 2 \sum_{i=1}^{n} (x_i - E(\underline{x}))(\bar{\underline{x}} - E(\underline{x})).$$

Nu geldt

$$\sum_{i=1}^{n} (x_i - E(\underline{x}))(\bar{\underline{x}} - E(\underline{x})) = (\bar{\underline{x}} - E(\underline{x}))(x_i - nE(\underline{x})) = n(\bar{\underline{x}} - E(\underline{x}))^2.$$

Hiervan gebruik makend vinden we

$$\sum_{i=1}^{n} (x_i - \bar{\underline{x}})^2 = \sum_{i=1}^{n} (x_i - E(\underline{x}))^2 - n(\bar{\underline{x}} - E(\underline{x}))^2 .$$

Beschouwen we nu de stochastiek (schatter van $\sigma^2(\underline{x})$)

$$\underline{s}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{x}_i - \bar{\underline{x}})^2$$

waarin $$\bar{\underline{x}} = \frac{1}{n} \sum_{i=1}^{n} \underline{x}_i$$

dan volgt hieruit

$$E(\underline{s}^2) = \frac{1}{n-1} E\{ \Sigma(\underline{x}_i - E(\underline{x}))^2 - n\, E(\bar{\underline{x}} - E(\underline{x}))^2\}$$

$$= \frac{1}{n-1} \{n\sigma^2 - n.\frac{\sigma^2}{n} \} = \sigma^2 .$$

De verwachtingswaarde van de steekproef variantie is dus gelijk aan de universum variantie. (Men zegt ook: de steekproef variantie is een zuivere schatting van de universum variantie.)

De vervanging van n door n - 1 is in practische toepassingen uiteraard alleen van belang voor kleine waarden van n.
Het getal n - 1 wordt het aantal vrijheidsgraden genoemd waarmee de steekproef variantie s^2 bepaald is.

13. De student-stochastiek t

De confidentie intervallen welke we in een voorafgaande sectie hebben afgeleid, veronderstellen de kennis van de universum spreiding $\sigma(\underline{x}) = \sigma$. Zo'n universum spreiding mag vaak bekend verondersteld worden bij routine metingen, dus als met stochastiek $\underline{x}$ reeds een grote ervaring is opgedaan. Is dit niet het geval dan maakt men als n niet te klein is ($n \geqslant 20$ zie onderstaande tabel) geen grote fout als in de formules van pag.55 de universum spreiding σ vervangen wordt door de steekproefspreiding s.
Voor kleine waarden van n biedt de mathematische statistiek echter een meer bevredigende oplossing. Als nl. $\underline{x}_i$, $i = 1,\ldots,n$ normale stochastieken zijn met dezelfde verwachtingswaarde $\mu = E(\underline{x}_i)$ en dezelfde spreiding $\sigma(\underline{x}_i) = \sigma$, dan is ook de schatter

$$\bar{\underline{x}} = \frac{1}{n} \sum_{i=1}^{n} \underline{x}_i$$

van de verwachtingswaarde $E(\underline{x})$ (zie pag.53) een normale stochastiek met verwachtingswaarde $\mu = E(\bar{\underline{x}})$ maar met spreiding $\sigma(\bar{\underline{x}}) = \frac{\sigma}{\sqrt{n}}$. (zie V_6 pagina 43)

De stochastische grootheid

$$\left(\frac{\frac{1}{n} \sum_{i=1}^{n} \underline{x}_i - \mu}{\sigma} \right) \sqrt{n} = \left(\frac{\bar{\underline{x}} - \mu}{\sigma} \right) \sqrt{n}$$

is dan een standaard normaal verdeelde grootheid.
Vervangt men nu het vaste getal σ door de stochastische grootheid $\underline{s} = \sqrt{\frac{1}{n-1} (\underline{x}_i - \frac{1}{n} \sum_{i=1}^{n} \underline{x}_i)^2}$ dan krijgt men de nieuwe stochastische variabele

$$\underline{t} = \frac{\frac{1}{n} \sum_{i=1}^{n} \underline{x}_i - \mu}{\underline{s}} \sqrt{n} = \left(\frac{\bar{\underline{x}} - \mu}{\underline{s}} \right) \sqrt{n},$$

waarvan men de distributie functie voor elke waarde van n kan berekenen; n-1 heet weer het aantal vrijheidsgraden van $\underline{t}$. $\underline{t}$ heet de student-stochastiek en de distributie functie $P(\underline{t} \leqslant t)$ de student verdeling.
Deze student verdeling is in de meeste statistische hand- en tabellenboeken getabelleerd.

Wil men nu op grond van een steekproef ter grootte n en met gebruikmaking van de student-verdeling een confidentie interval voor $\mu = E(\underline{x})$ bepalen, behorende bij een onbetrouwbaarheid α dan zoeke men in deze tabel, voor n-1 vrijheidsgraden de waarde t_α zodat

$$P(|\underline{t}| \geq t_\alpha) = \alpha.$$

De formule drukt uit dat bij realisatie van $\bar{\underline{x}}$ en van $\underline{s}$, dus bij het nemen van een steekproef ter grootte n, de kans op de gebeurtenis

$$\bar{x} - t_\alpha \frac{s}{\sqrt{n}} \leq E(\underline{x}) \leq \bar{x} + t_\alpha \frac{s}{\sqrt{n}}$$

gelijk aan α is.

Het interval

$$\left(\bar{x} - t_\alpha \frac{s}{\sqrt{n}}, \bar{x} + t_\alpha \frac{s}{\sqrt{n}}\right)$$

wordt weer een confidentie interval voor de verwachtingswaarde $E(\underline{x})$ genoemd behorende bij de onbetrouwbaarheid α.

De volgende tabel geeft een vergelijking van de kansen

$$P\left(\left|\frac{\bar{\underline{x}} - \mu}{\sigma}\sqrt{n}\right| \geq t\right)$$ afgeleid uit de standaard normale verdeling

en

$$P\left(\left|\frac{\bar{\underline{x}} - \mu}{\underline{s}}\sqrt{n}\right| \geq t\right)$$ afgeleid uit de student verdeling.

Zij toont aan dat voor de waarden $n \geq 20$ de student verdeling en de normale verdeling tot practisch gelijke confidentie intervallen leiden.

	$P\left(\left\|\frac{\bar{\underline{x}} - \mu}{\sigma}\sqrt{n}\right\| \geq t\right)$	$P\left(\left\|\frac{\bar{\underline{x}} - \mu}{\underline{s}}\sqrt{n}\right\| \geq t\right)$				
t	Normale verdeling	Student verdeling				
		n= 3	5	8	10	20
1	0.32	0.38	0.35	0.33	0.33	0.32
1.5	0.14	0.22	0.20	0.18	0.15	0.14
2	0.05	0.15	0.10	0.08	0.07	0.05
2.5	0.008	0.11	0.05	0.04	0.04	0.02
3	0.003	0.06	0.04	0.018	0.01	0.008
3.5	0.001	0.04	0.018	0.011	0.007	0.003
4	0.000	0.03	0.01	0.007	0.005	0.000

14. Interpretatie van meetresultaten

We hebben gezien dat het meten van een grootheid X neerkomt op het nemen van een steekproef, d.w.z. uit het enkele keren herhaalde realiseren van de stochastische variabele $\underline{x}$, die met deze meting is geassocieerd.
Laten we veronderstellen dat:
$\underline{x}$ een normale stochastiek is (zie centrale limietstelling, pag 40) en dat de steekproefuitkomsten $x_1,\ldots,x_n$ onafhankelijk van elkaar bepaald zijn.

Zo'n steekproef levert dan de volgende informatie:

1. de beste schatting die we hebben van de werkelijke waarde X is het rekenkundig gemiddelde $\bar{x}$ van de steekproefuitkomsten:

$$X \sim \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i .$$

2. een maat voor de onnauwkeurigheid van deze schatting $\bar{x}$ van X is de steekproefspreiding $s(\underline{\bar{x}})$ van $\underline{\bar{x}}$, waarbij

$$s(\underline{\bar{x}}) = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n(n-1)}} .$$

Men ziet onmiddellijk dat $s(\underline{\bar{x}}) = \frac{s}{\sqrt{n}}$, als s de steekproef van de stochastiek $\underline{x}$ is; zie pag. 56). Kent men op grond van ervaring de werkelijke spreiding σ van $\underline{x}$ (hetgeen bij routine metingen vaak het geval is) dan is

$$\sigma(\underline{\bar{x}}) = \frac{\sigma}{\sqrt{n}}$$

een betere maat voor de onnauwkeurigheid van de schatting $\bar{x}$ van X.

Bij het vermelden van meetresultaten in rapporten of artikelen zal men in het algemeen niet alle steekproefuitkomsten opgeven. Voor een juiste beoordeling van de gevonden resultaten is het echter nodig naast het gemiddelde meetresultaat $\bar{x}$, als beste schatting van datgene wat men heeft willen meten, ook een schatting op te geven voor de onnauwkeurigheid van dit gemiddelde. Deze schatting is ofwel $\sigma(\underline{\bar{x}})$ ofwel $s(\underline{\bar{x}})$ waarbij in dit laatste geval ook de steekproef-

grootte moet worden vermeld.
De informatie die uit deze gegevens is af te leiden is
a) bij gegeven onbetrouwbaarheid α geldt voor X het confidentie interval

$$\bar{x} - t_\alpha s(\bar{x}) \leq X \leq \bar{x} + t_\alpha s(\bar{x})$$

waarbij t_α wordt afgelezen uit een tabel voor de student verdeling met n-1 graden van vrijheid en zo dat

$$P(|\underline{t}| \geq t_\alpha) = \alpha.$$

Is $\sigma(\bar{x})$ gegeven dan wordt het confidentie interval

$$\bar{x} - u_\alpha \sigma(\bar{x}) \leq X \leq \bar{x} + u_\alpha \sigma(\bar{x})$$

waarbij u_α wordt afgelezen uit een tabel voor de standaard normale verdeling en wel zo dat

$$P(|\underline{u}| \geq u_\alpha) = \alpha .$$

Een veel voorkomende waarde van α is 0.05. In het geval dat $\sigma(\bar{x})$ bekend is wordt dan het confidentie interval

$$\bar{x} - 2\sigma(\bar{x}) \leq X \leq \bar{x} + 2\sigma(\bar{x})$$

b. Wil men nagaan of een gegeven getal Y niet aanvaardbaar is als werkelijke waarde van de te meten grootheid dan kieze men eerst een kans α welke men wil accepteren op onjuistheid van de volgende uitspraak:

het getal Y is niet aanvaardbaar als werkelijke waarde van de te meten grootheid als Y ligt buiten het bij de onbetrouwbaarheid α behorende confidentie interval;

Men kan ook zeggen: als Y binnen dit confidentie interval ligt, dan kan op grond van de steekproefuitkomsten niet beweerd worden dat de werkelijke waarde van de te meten grootheid verschilt van Y.

15. Foutenvoortplanting en foutendiscussie

Het komt vaak voor dat men een grootheid, welke men wenst te kennen, niet rechtstreeks kan meten. Wil men bijv. de versnelling g van de zwaartekracht bepalen met behulp van de slinger, dan meet men de slingertijd T en de slingerlengte l, waarna g berekend wordt uit de formule

$$g = 4\pi^2 \frac{l}{T^2} .$$

T en l zijn grootheden die gemeten moeten worden. Men kent dus alleen schattingen $\hat{T}$ en $\hat{l}$ van hun werkelijke waarden. Bovendien kent men de steekproef spreidingen $s_{\hat{T}}$ en $s_{\hat{l}}$ welke de onnauwkeurigheid representeren waarmee T en l door $\hat{T}$ en $\hat{l}$ geschat worden.

Voor de hand liggende vragen zijn nu:

1. is $\hat{g} = 4\pi^2 \frac{\hat{l}}{\hat{T}^2}$ een goede schatting van g ?

2. hoe werkt de onnauwkeurigheid in de schatting van T en l door in de onnauwkeurigheid van de schatting van g ?

Voor een algemene behandeling van dit probleem veronderstellen we dat Y een bekende functie is van de meetbare grootheden X_i, i = 1,...,m:

$$Y = f(X_1, X_2, \ldots, X_m).$$

Als aan de grootheid X_i de stochastiek $\underline{x}_i$ is geassocieerd, dan is aan Y de stochastiek

$$\underline{y} = f(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_m)$$

geassocieerd.

De stochastieken $\underline{x}_i$ worden onafhankelijk van elkaar verondersteld, terwijl ter verkorting van de notatie gesteld wordt:

$$E(\underline{x}_i) = \mu_i \quad \text{en} \quad \sigma(\underline{x}_i) = \sigma_i.$$

Eerste geval: Y is een lineaire functie van de grootheden X_i;

$$Y = \sum_{i=1}^{m} a_i X_i.$$

Zijn de $\underline{x}_i$ normale stochastieken (vaak aanneemelijk; zie centrale limietstelling pag 40), dan is ook $\underline{y}$ een normale stochastiek (zie bewering pag 55) terwijl geldt:

$$E(\underline{y}) = \sum_{i=1}^{m} a_i E(\underline{x}_i) = \sum_{i=1}^{m} a_i \mu_i$$

en

$$\sigma^2(\underline{y}) = \sum_{i=1}^{m} a_i^2 \sigma^2(x_i) = \sum_{i=1}^{m} a_i^2 \sigma_i^2$$

(eigenschap V_4 pag 42 ; denk aan veronderstelde onafhankelijkheid van de stochastieken $\underline{x}_i$!). Is de onafhankelijkheidsveronderstelling niet gerechtvaardigd dan moet voor $\sigma^2(\underline{y})$ de volgende formule gebruikt worden:

$$\sigma^2(\underline{y}) = \sum_{i=1}^{m} a_i^2 \sigma_i^2 + \sum_{i \neq j} a_i a_j \operatorname{cov}(x_i, x_j).$$

(zie V_4 pag 43)

Is de aanname "$\underline{x}_i$ normaal" niet voor alle i gerechtvaardigd, maar zijn de stochastieken $\underline{x}_i$ wel onafhankelijk van elkaar, dan mag de stochastiek $\underline{y}$ wegens de centrale limietstelling, meestal toch normaal verondersteld worden met bovenstaande verwachtingswaarde en variantie.

De spreiding σ_i wordt ook wel de absolute fout genoemd die optreedt bij meting van de grootheid X_i (eigenlijk is σ_i een maat voor de mogelijke fout die bij deze vervanging van X_i door de steekproefwaarde x_i kan optreden).
De naam "absolute fout" staat tegenover "relatieve fout", waarvoor men als maat neemt:

$$\frac{\sigma_i}{\mu_i} .$$

In plaats van relatieve fout gebruikt men ook de naam: variatie coefficient.
$\sigma_{\underline{Y}}$ is een maat voor de absolute fout in Y, terwijl, als Y lineair met de te meten grootheden X_i samenhangt, de volgende "wet van de fouten voortplanting" blijkt te gelden:

$$\sigma_{\underline{Y}}^2 = \sum_{i=1}^{m} a_i^2 \sigma_i^2 .$$

Tweede geval: Y is een niet lineaire functie van de grootheden X_i, $i = 1,\ldots,m$:

$$Y = f(X_1, X_2, \ldots, X_m).$$

In plaats van de stochastiek $\underline{x}_i$ gebruiken we de stochastiek $\Delta\underline{x}_i$ zo gedefinieerd dat $\underline{x}_i = E(\underline{x}_i) + \Delta\underline{x}_i$. Dan geldt

$$E(\Delta\underline{x}_i) = 0 \quad \text{en} \quad \sigma(\Delta\underline{x}_i) = \sigma_i .$$

Veronderstellen we nu nog dat de spreidingen σ_i klein zijn, dus dat de kans op grote waarden voor Δx_i klein is, dan levert Taylorontwikkeling van $f(X_1, X_2, \ldots, X_m)$ en verwaarlozing van alle termen van hogere dan de eerste graad:

$$\underline{Y} = f(E(\underline{x}_1), E(\underline{x}_2), \ldots, E(\underline{x}_m)) + \sum_{i=1}^{m} \left(\frac{\partial f}{\partial X_i}\right) . \Delta\underline{x}_i .$$

Ook nu geldt weer, wegens de centrale limietstelling, dat de stochastiek $\underline{Y}$ meestal normaal verondersteld mag worden, als tenminste de stochastieken $\Delta\underline{x}_i$ onafhankelijk van elkaar zijn.
Hieruit berekenen we

$$E(\underline{y}) = f(E(\underline{x}_1), E(\underline{x}_2), \ldots, E(\underline{x}_m))$$

en

$$\sigma^2(\underline{y}) = \sum_{i=1}^{m} \left(\frac{\partial f}{\partial X_i}\right)^2_{E(\underline{x}_i)} \sigma_i^2 .$$

Interpreteren we weer $\sigma_{\underline{y}}$ als maat voor de absolute fout bij de meting van X_i dan luidt de algemene fouten voortplantingswet:

$$\sigma_{\underline{y}}^2 = \sum_{i=1}^{m} \left(\frac{\partial f}{\partial X_i}\right)^2_{E(\underline{x}_i)} \sigma_i^2 .$$

(Men lette er op dat deze wet is afgeleid onder de veronderstelling dat de stochastieken $\underline{x}_i$, $i = 1,\ldots,m$ onderling onafhankelijk zijn en dat de spreidingen σ_i klein zijn!)

Deze algemene wet van de fouten voortplanting krijgt een speciale vorm als de grootheid Y op een zuiver multiplicatieve wijze samenhangt met de grootheden X_i, $i = 1,\ldots,m$:

$$Y = X_1^{\alpha_1} . X_2^{\alpha_2} \ldots\ldots . X_m^{\alpha_m} .$$

Men vindt dan als fouten voortplantingswet:

$$\left(\frac{\sigma_{\underline{y}}}{\mu_{\underline{y}}}\right)^2 = \sum_{i=1}^{m} \alpha_i^2 \left(\frac{\sigma_i}{\mu_i}\right)^2 .$$

In dit geval zijn de relatieve fouten in de metingen van X_i dus van meer belang dan de absolute fouten in deze metingen.

16. De methode van de kleinste kwadraten

In dit hoofdstuk beschouwen we situaties welke gekenmerkt zijn door een meetbare grootheid Y en een stelsel meetbare grootheden $X_1,\ldots,X_m$. Het onderscheid tussen Y en $X_1,\ldots,X_m$ bestaat hierin dat we Y zullen zien als een grootheid welke door de grootheden $X_1,\ldots,X_m$ (min of meer) bepaald wordt. Denk bv. aan een chemisch proces met Y als opbrengst en met $X_1,\ldots,X_m$ als proces parameters: procestemperatuur, invoer concentraties van met elkaar reagerende stoffen, enz.

Het is voor onze beschouwingen niet nodig dat Y functioneel met $X_1,\ldots,X_m$ samenhangt dus dat een relatie geldt van de vorm:

$$Y = f(X_1,\ldots,X_m)$$

Er kunnen nl. nog andere grootheden zijn, welke we niet "onder controle" hebben of zelfs niet kennen, maar welke toch invloed hebben op Y.

De vraag waar we ons in het bijzonder mee bezighouden is: door welke lineaire relatie tussen Y en $X_1,\ldots,X_m$ wordt een eventueel verband tussen deze grootheden "het beste" benaderd.

Zij Z een m+1 vector met componenten $Z_o = Y$ en $Z_i = X_i$, $i = 1,\ldots,m$.
Met meting van Z is weer een stochastiek $\underline{z}$ geassocieerd.
$\underline{z}$ is een stochastische vector met stochastische componenten $\underline{z}_o = \underline{y}$ en $\underline{z}_i = \underline{x}_i$, $i = 1,\ldots,m$. Uiteraard zijn de stochastieken $\underline{y}$ en $\underline{x}_i$ door het meet mechanisme geassocieerd aan Y en X_i.

Laat nu $z^1,\ldots,z^k$ een steekproef ter grootte k zijn. Dit betekent dat de volgende k rijtjes van getallen bekend zijn:

y_1	x_{11}	x_{12}	. . .	x_{1m}
y_2	x_{21}	x_{22}	. . .	x_{2m}
.	.	.		.
.	.	.		.
.	.	.		.
y_k	x_{k1}	x_{k2}	. . .	x_{km}.

Onder de beste lineaire benadering van een eventuele relatie tussen Y en $X_1,\ldots,X_m$ welke op grond van de steekproef ter grootte k berekend kan worden verstaan we nu <u>per definitie</u> de relatie

$$Y = \hat{a}_o + \hat{a}_1 X_1 + \ldots + \hat{a}_m X_m$$

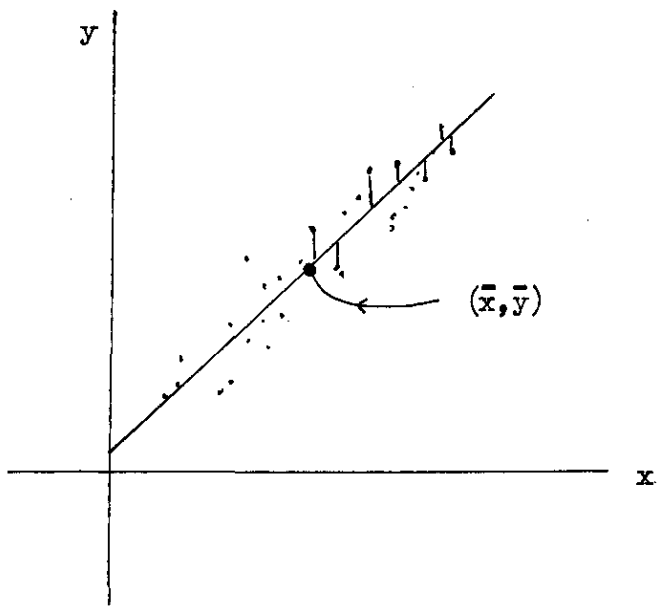
waarbij de kwadratische vorm

$$\varphi(a_o,a_1,\ldots,a_m) = \sum_{j=1}^{k} (a_o + \sum_{i=1}^{m} a_j x_{ij} - y_i)^2$$

haar minimale waarde aanneemt in $(\hat{a}_o,\hat{a}_1,\ldots,\hat{a}_m)$. De aldus gedefinieerde beste lineaire benadering van het verband tussen Y en $X_1,\ldots,X_m$ noemt men een "kleinste kwadraten benadering; de methode volgens welke deze schatting verkregen wordt, wordt een "kleinste kwadraten methode" genoemd.

We zullen ons beperken tot het geval m = 1 dus tot situaties welke gekenmerkt worden door één grootheid Y en één grootheid X.

De steekproef bestaat dan uit k getallen paren (y_j,x_j), $j = 1,\ldots,k$. Deze getallen paren representeren een puntenwolk in een plat vlak. Het probleem is een rechte lijn $y = \hat{b} + \hat{a}x$ te trekken door deze puntenwolk, zodat de som van de kwadraten van de "afstanden" van deze punten tot die rechte, gemeten evenwijdig aan de y-as, minimaal is.



De te minimaliseren functie is nu

$$\varphi(a,b) = \sum_{j=1}^{k} (y_j - b - ax_j)^2.$$

De nodige (en voldoende voorwaarde) voor minimaliteit is het nul zijn van de partiële afgeleiden van $\varphi(a,b)$ naar a en b:

$$\frac{\partial\varphi(a,b)}{\partial a} = 0 \quad \Rightarrow \quad \hat{a}\sum_{j=1}^{k} x_j^2 + \hat{b}\sum_{j=1}^{k} x_j = \sum_{j=1}^{k} x_j y_j$$

$$\frac{\partial\varphi(a,b)}{\partial b} = 0 \qquad \hat{a}\sum_{j=1}^{k} x_j + \hat{b}\,k + \sum_{j=1}^{k} y_j$$

Hieruit volgt *) :

$$\hat{a} = \frac{k\,\Sigma x_j y_j - \Sigma x_j\,\Sigma y_j}{k\,\Sigma x_j^2 - (\Sigma x_j)^2}$$

$$\hat{b} = \frac{\Sigma x_j\,\Sigma x_j y_j - \Sigma y_j\,\Sigma x_j^2}{k\,\Sigma x_j^2 - (\Sigma x_j)^2} .$$

We kunnen deze formules nog iets beter hanteerbaar maken.
Uit de bovenstaande relatie

$$\hat{a}\sum_{j=1}^{k} x_j + \hat{b}\,k = \sum_{j=1}^{k} y_j ,$$

volgt, n.l. na deling door k en stellend

$$\bar{x} = \frac{1}{k}\,\Sigma x_j \qquad \bar{y} = \frac{1}{k}\sum_{j=1}^{k} y_j ,$$

dat

$$\hat{a}\bar{x} + \hat{b} = \bar{y}.$$

Het punt $(\bar{x},\bar{y})$ is het zwaartepunt van de puntenwolk. De best passende rechte lijn gaat dus door het zwaartepunt van de puntenwolk.

---

*) Om de formules niet nodeloos ingewikkeld te maken zijn de sommatiegrenzen niet altijd expliciet vermeld.
Deze grenzen zijn uit de tekst echter direkt duidelijk.

Nemen we nu $\bar{x}$ op de x-as tot oorsprong, dan moet in de bovenstaande formules x_j door nul vervangen worden, hetgeen leidt tot

$$\hat{a} = \frac{\Sigma x_j y_j}{\Sigma x_j^2},$$

en

$$\hat{b} = \frac{1}{k} \Sigma y_j.$$

De hier behandelde methode van de kleinste kwadraten is toe te passen op elke twee dimensionale puntenwolk. Ze is ook direct uit te breiden tot puntenwolken in ruimten van hogere dimensies. Aan de gevonden schattingen hebben we echter nog geen statistische consequenties verbonden. Willen we dit wel doen, dan moeten we onze veronderstellingen gaan verfijnen.

We nemen nu aan:

1) Er bestaat een (echte) lineaire relatie tussen de grootheid Y en de grootheid X.

   $$Y = \beta + \alpha X$$

   X wordt de onafhankelijke variabele genoemd,
   Y wordt de afhankelijke variabele genoemd.
2) X kan foutloos gemeten worden; d.w.z. de aan X geassocieerde stochastische variabele $\underline{x}$ heeft een spreiding 0.
3) De onnauwkeurigheid bij de meting van Y is onafhankelijk van de waarde die X tijdens de meting aanneemt. De spreiding van de aan Y geassocieerde stochastiek $\underline{y}_i$ is σ.
4) De steekproefgrootte is k; het nemen van de steekproef behelst het successievelijk toekennen van k waarden x_j, $(j = 1,\ldots,k)$ aan X en het meten van de bijbehorende waarde y_j van Y. Deze k metingen van Y worden onderling onafhankelijk verondersteld.
   De steekproefuitkomsten zijn (y_j,x_j), $j = 1,\ldots,k$.
   De oorsprong voor de variabele x is zodanig gekozen dat $\sum_{j=1}^{k} x_j = 0$.

Uit het voorgaande volgt nu: de schatting $\hat{a}$ van α, die we op grond van de steekproef en met behulp van de methode van de kleinste kwadraten krijgen kan beschouwd worden als een realisatie van de stochastiek

$$\underline{a} = \frac{1}{\Sigma x_j^2} \Sigma x_j \underline{y}_j$$

De schatting $\hat{b}$ van β, welke we op deze wijze verkrijgen kan beschouwd worden als een realisatie van de stochastiek

$$\underline{\hat{b}} = \frac{1}{k} \Sigma \underline{y}_j .$$

Wegens de onafhankelijkheid van de steekproefresultaten, levert toepassing van de centrale limietstelling (zie pag 40) dat $\underline{\hat{a}}$ en $\underline{\hat{b}}$ bij benadering normale stochastieken zijn.

De verwachtingswaarde $E(\underline{\hat{a}})$ is gelijk aan de werkelijke waarde α.

$$E(\underline{\hat{a}}) = \frac{1}{\Sigma x_j^2} \Sigma x_j E(\underline{y}_j)$$

$$= \frac{1}{\Sigma x_j^2} \Sigma x_j (\beta + \alpha x_j)$$

$$= \frac{\beta}{\Sigma x_j^2} \Sigma x_j + \alpha \frac{\Sigma x_j^2}{\Sigma x_j^2} = \alpha .$$

De verwachtingswaarde van $E(\underline{b})$ is gelijk aan de werkelijke waarde b.

$$E(\underline{\hat{b}}) = \frac{1}{k} \Sigma E(\underline{y}_j)$$

$$= \frac{1}{k} \Sigma (\beta + \alpha x_j)$$

$$= \beta + \frac{1}{k} \alpha \Sigma x_j = \beta .$$

Deze eigenschap drukt men uit met de woorden:
$\hat{a}$ en $\hat{b}$ zijn "zuivere" schattingen van α en β.

De stochastiek $\hat{\underline{a}}$ heeft een spreiding $\sigma_{\hat{\underline{a}}} = \frac{\sigma}{\sqrt{\Sigma x_j^2}}$.

De stochastiek $\hat{\underline{b}}$ heeft een spreiding $\sigma_{\hat{\underline{b}}} = \frac{\sigma}{\sqrt{k}}$.

Kent men σ, dus kent men de onnauwkeurigheid in de meting van Y, dan kan bij gegeven onbetrouwbaarheid γ een confidentie interval aangegeven worden zowel voor α als voor β.

$$\frac{1}{\Sigma x_j^2}\left(\Sigma x_j y_j - u_\gamma \sigma\sqrt{\Sigma x_j^2}\right) \leq \alpha \leq \frac{1}{\sqrt{x_j^2}}\left(\Sigma x_j y_j + u_\gamma \sigma\sqrt{\Sigma x_j^2}\right)$$

$$\frac{1}{k}(\Sigma y_j - u_\gamma \sigma\sqrt{k}) \leq \beta \leq \frac{1}{k}(\Sigma y_j + u_\gamma \sigma\sqrt{k}),$$

waarbij u_γ zo gekozen wordt uit een tabel voor de standaard normale verdeling dat

$$P(|u| \geq u_\gamma) = \gamma.$$

Is de onnauwkeurigheid in de meting van Y niet vooraf bekend, dus is σ niet gegeven, dan levert de steekproef hiervoor een schatting s, hetgeen als volgt blijkt. De functie $\varphi(a,b) = \Sigma_j (y_j - ax_j - b)^2$ neemt haar minimale waarde aan voor $a = \hat{a}$ en $b = \hat{b}$ waarbij (zie pag 67) $\hat{a}$ en $\hat{b}$ voldoen aan

$$\Sigma x_j(y_j - \hat{a}x_j - \hat{b}) = 0$$

$$\Sigma (y_j - \hat{a}x_j - \hat{b}) = 0$$

terwijl de werkelijke waarde $Y_j = E(\underline{y}_j)$ voldoet aan

$$\Sigma Y_j - \alpha x_j - \beta = 0$$

(bedenk dat $\underline{x}_j$ foutloos is zodat x_j gelijk is aan de werkelijke waarde X_i).

Hieruit volgt

$$\varphi(\hat{\underline{a}},\hat{\underline{b}}) = \Sigma\,(\underline{y}_j - \hat{\underline{a}}x_j - \hat{\underline{b}})^2$$

$$= \Sigma\,((\underline{y}_j - Y_j) - (\hat{\underline{a}} - \alpha)x_j - (\hat{\underline{b}} - \beta))^2$$

$$= \Sigma\,(\underline{y}_j - Y_j)^2 + (\hat{\underline{a}} - \alpha)^2 \Sigma x_j^2 + k(\hat{\underline{b}} - \beta)^2 +$$

$$- 2(\hat{\underline{a}} - \alpha)\ \Sigma x_j((\underline{y}_j - Y_j) - (\hat{\underline{b}} - \beta)) +$$

$$- 2\,(\hat{\underline{b}} - \beta)\ \Sigma\,(\underline{y}_j - Y_j),$$

waaruit met behulp van bovenstaande relaties volgt:

$$\varphi(\hat{\underline{a}},\hat{\underline{b}}) = \Sigma\,(\underline{y}_j - Y_j)^2 - (\hat{\underline{a}} - \alpha)^2\ \Sigma\, x_j^2 - k(\hat{\underline{b}} - \beta)^2.$$

Nemen we de verwachtingswaarde van $\varphi(\hat{\underline{a}},\hat{\underline{b}})$, dan vinden we

$$E\left(\varphi(\hat{\underline{a}},\hat{\underline{b}})\right) = \Sigma\,E(\underline{y}_j - Y_j)^2 - \Sigma x_j^2\ E(\hat{\underline{a}} - \alpha)^2 - k\,E(\hat{\underline{b}} - \beta)^2$$

$$= k\,\sigma^2 - \sigma^2 - \sigma^2$$

$$= (k - 2)\,\sigma^2.$$

De grootheid $s^2 = \dfrac{\varphi(\hat{a},\hat{b})}{k - 2} = \dfrac{\Sigma\,(y_j - \hat{a}x_j - \hat{b})^2}{k - 2}$ is dus een (zuivere) schatting van de variantie van σ^2.

$k - 2$ wordt het aantal vrijheidsgraden genoemd waarmee s bepaald wordt.

Ook nu is weer bij gegeven onbetrouwbaarheid γ een confidentie interval voor α en β te berekenen:

$$\frac{1}{\Sigma\, x_j^2}\,(\Sigma x_j y_j - t_\gamma s\sqrt{\Sigma x_j^2}) \leq \alpha \leq \frac{1}{\Sigma\, x_j^2}\,(\Sigma x_j y_j + t_\gamma s\sqrt{\Sigma\, x_j^2}\,)$$

$$\frac{1}{k}\left(\Sigma y_j - t_\gamma s^2 \sqrt{k}\right) < \beta < \frac{1}{k}\left(\Sigma y_j + t_\gamma s \sqrt{k}\right)$$

waarbij t_γ wordt opgezocht in een tabel voor de student verdeling, behorende bij $k - 2$ graden van vrijheid en zo dat

$$P(|t| \geq t_\gamma) = \gamma.$$

Een soortgelijke methode kan ook worden gebruikt in geval de grootheid Y lineair samenhangt met m grootheden X_i, $i = 1,\ldots,m$ en men de coefficienten van deze lineaire relatie wil schatten.
Het rekenwerk wordt dan veel meer ingewikkeld. Rekenmachine programma's zijn echter aanwezig welke in vele voorkomende gevallen zonder meer kunnen worden toegepast.