Kansrekening en Statistiek
J. de Ruiter
Februari 2022

Inhoud

0. Inleiding

We leggen de volgende vijf problemen voor.

a. Het driedeurenprobleem

Een probleem dat publieke bekendheid kreeg door spelshows op tv, o.a. van Monty Hall in Amerika en Willem Ruis in Nederland.
Het probleem komt op het volgende neer:

De quizwinnaar mag kiezen uit drie deuren. Achter een van de drie deuren zit een grote prijs, achter de andere twee deuren zit niks. Als de deelnemer nu een van de drie deuren aanwijst (maar nog niet geopend heeft), opent de quizmaster (die weet waar de prijs zit) een deur waarachter de prijs niet zit. De quizmaster biedt de deelnemer de gelegenheid alsnog de overblijvende deur te kiezen.
Wat moet de deelnemer nu doen?

De meeste mensen denken intuïtief dat dit niets uitmaakt. De quizmaster opent een deur waar geen prijs achter zit. De prijs zit dus achter een van de twee andere deuren, waarvan een al aangewezen is door de deelnemer. Wat maakt het dan uit of je nog wisselt? Echter, als je wel wisselt, dan neemt de kans op de prijs toe van 1/3 naar 2/3!
Voor veel mensen blijft het moeilijk om te geloven dat de intuïtieve redenering fout is, ook na de nodige uitleg.
Voor uitleg: zie par. 2.

b. Het verjaardagprobleem
In een zaal bevinden zich n personen.
Hoe groot is de kans dat minstens twee personen op dezelfde dag jarig zijn?
Wat blijkt? Bij n = 23 is deze kans al meer dan 50%. Bij n = 41 zelfs al meer dan 90%!
Dit probleem zal in par. 2 worden behandeld.

c. Het krentenprobleem
Een bakker deed eens het volgende experiment: hij maakte deeg voor 100 broodjes en mengde daarin precies 100 krenten. Na nauwkeurig mengen werden de broodjes gebakken.
Hoeveel broodjes zullen nu 0 krenten bevatten?
Wat bleek? 37 broodjes bevatten 0 krenten!
Wat zegt de kansrekening nu m.b.t. dit experiment?
Dit zal in par. 2 uit de doeken worden gedaan.

d. Testen op besmetting
Van een bepaald virus is bekend dat 1% van de populatie ermee besmet is. Van een bepaalde test weten we dat 95% van de besmetten positief reageert en 6% van de onbesmetten.
Als iemand positief reageert, wat is dan de kans dat er werkelijk sprake is van besmetting?
Het anwoord is: deze kans is 13,8%!
In par. 1 zullen we zien hoe deze kans, een zgn. voorwaardelijke kans, berekend kan worden.

e. Zero knowledge
A heeft twee ballen, een rode en een groene. B is kleurenblind en wil niet geloven dat de kleuren verschillend zijn. Een methode om B te overtuigen is de volgende:
B krijgt beide ballen in zijn handen. A kijkt niet en B mag dan naar keuze beide ballen wel of niet verwisselen. Daarna mag A weer kijken om te vertellen of B wel of niet de ballen heeft verwisseld. Dit wordt verscheidene keren herhaald en elk keer moet A weer zeggen of de ballen verwisseld zijn of niet.

Als beide ballen dezelfde kleur zouden hebben, dan zou A elke keer moeten raden of er wel of niet verwisseld is. Dat gaat A niet elke keer lukken. Maar het gaat A wel elke keer lukken. Dus B moet concluderen dat de ballen verschillend van kleur zijn.

Bij zogeheten 'zero knowledge-bewijzen gaat het om technieken om te bewijzen dat je kennis hebt van bepaalde geheime informatie, zónder die informatie openbaar te maken. Ze zijn onmisbaar in de bankenwereld: als je de bank bezoekt – aan het loket of digitaal – moet de bank kunnen verifiëren wie je bent. Je toetst een code in, maar om veiligheidsredenen heeft de bank jouw persoonlijke code niet opgeslagen. Kan de bank een code controleren die hij zelf niet kent? Ja, is het antwoord, dankzij 'zero knowledge.
Voor meer informatie over zero knowledge: zie bijv. NRC 19 maart 2021.

De vijf voorgaande voorbeelden maken duidelijk dat kans een begrip is dat vaak schuurt met onze intuïtie, maar dat er ook veel toepassingen zijn waarmee concrete problemen opgelost kunnen worden.
Kansrekening heeft een lange weg doorlopen in de geschiedenis van de wiskunde. Uiteindelijk is dit een serieus onderdeel van de zuivere en toegepaste wiskunde geworden.
In dit document is getracht een inleiding in dit onderdeel van de wiskunde te geven waarbij:

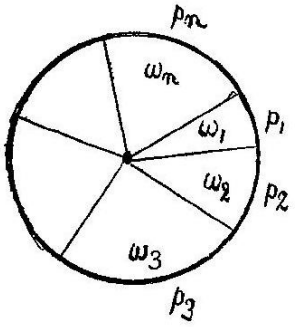
- de onderliggende wiskunde correct is,
- de probleemstellingen heel concreet zijn,
- de lezer de belangrijkste hoofdstukken uit de kansrekening en statistiek voorgeschoteld krijgt,
- de lezer niet onnodig abstracte begrippen voorgelegd krijgt die weliswaar bij verdere studie van kansrekening en statistiek relevant zijn, maar in eerste instantie nog niet direct een rol spelen,
- meer specialistische statische methoden niet behandeld worden.

Toevalsexperimenten kunnen eindig veel, aftelbaar oneindig veel of overaftelbaar veel mogelijke uitkomsten hebben. De ontwikkeling van de kansrekening is begonnen met het bestuderen van toevalsexperimenten die slechts eindig veel mogelijke uitkomsten hebben. Pas later is de theorie verder uitgebreid voor toevalsexperimenten met oneinig veel uitkomsten. Met name de stap naar overaftelbaar veel mogelijke uitkomsten bleek een stevige theoretische uitdaging te zijn. Bijv. omdat hier niet meer een kans aan afzonderlijke uitkomsten toegekend kan worden, maar alleen aan bepaalde verzamelingen van uitkomsten (gebeurtenissen genoemd).
In de laatste paragrafen zullen we ons beperken tot die delen theorie die betrekking hebben op toevalsexperimenten met oneindig veel mogelijke uitkomsten, maar alleen voor zover het om concrete en realistische toepassingen gaat en zullen we verdergaande en abstractere onderdelen van de theorie vermijden.

Kansrekening: het berekenen van kansen.
Statistiek: het schatten van kansen uit waarnemingen.

1. Toevalsexperimenten met een eindig aantal uitkomsten

Veel (zo niet alle) toevalsexperimenten met een eindig aantal mogelijke uitkomsten zijn in feite te beschouwen als het draaien aan een geluksrad:



Voorbeelden
Werpen met een zuivere dobbelsteen; dat komt neer op het draaien aan een geluksrad met 6 sectoren van 60^o.
Aselect trekken van een toevalscijfer 0, 1, 2, ……of 9, ofwel het draaien aan een geluksrad met 10 sectoren van 36^0.
Het werpen met een zuivere munt. Mogelijke uitkomsten: kruis en munt.

De uitkomstenverzameling geven we aan met $\Omega = \{\omega_1, \ldots..,\omega_n\}$.
Als de cirkel omtrek 1 heeft, dan worden de booglenten $p_1, \ldots.., p_n$ van de n sectoren de kansen van de uitkomsten genoemd.
Dus $p(\omega_i) > 0$ en $\sum p(\omega_i) = 1$.
Een gebeurtenis A is een deelverzameling van bepaalde uitkomsten, dus $A \subset \Omega$.
De kans op de gebeurtenis A wordt gedefiniëerd als: $P(A) = \sum_{\omega_i \epsilon A} p(\omega_i)$.
Ook wordt gedefiniëerd: $P(\emptyset) = 0$.

Op deze manier is er een functie P ontstaan die aan elke deelverzameling A van Ω een getal toekent met de eigenschappen:

1. $P(A) \geq 0$
2. $P(\Omega) = 1$
3. P is additief d.w.z. $P(A \cup B = P(A) + P(B)$ als $A \cap B = \emptyset$

Zo'n functie heet kansfunctie.

Stelling

$$P(A) = \frac{\text{het aantal gunstige uitkomsten}}{\text{het aantal mogelijke uitkomsten}}$$

onder de voorwaarde dat de mogelijke uitkomsten van het experiment onderling fysisch gelijkwaardig zijn.

De fysische gelijkwaardigheid betekent dus dat $p(\omega_i) = 1/n$ voor i = 1 t/m n. Men noemt de kansfunctie dan symmetrisch. Het rad van avontuur is dan radiaal-symmetrisch.

Deze karakterisering van P(A) wordt ook vaak gebruik als definitie van kans (definitie van Laplace (1812).

Somregel voor kansen:
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Bewijs:
$P(A \cup B) = \sum_{\omega \in A \cup B} p(\omega) = \sum_{\omega \in A} p(\omega) + \sum_{\omega \in B} p(\omega) - \sum_{\omega \in A \cap B} p(\omega)$

Als $A \cap B = \emptyset$, dan zegt men dat de gebeurtenissen A en B elkaar uitsluiten of onverenigbaar zijn. In dat geval geldt dus $P(A \cup B) = P(A) + P(B)$.

Complementregel voor kansen:
$P(\neg A) = 1 - P(A)$

Bewijs volgt direct uit de somregel.

Voorbeeld
Eerst wordt geworpen met een zuivere dobbelsteen. Als s de ogensom is, dan volgt daarna nog een worp met s zuivere munten. Bereken de kans dat daarbij een “kruis” zit.
Als u dit correct berekent, dan vindt u als antwoord 107/128.

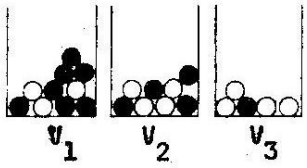
Stelling
Als A_1 t/m A_n elkaar uitsluiten, dan geldt $P(A_1 \cup ..... \cup A_n) = P(A_1) + ... + P(A_n)$.

De voorwaardelijke kans op de gebeurtenis A, gegeven B, is gedefiniëerd als:

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

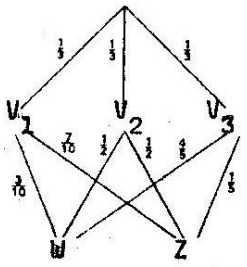
Ook genoteerd als $P_B(A)$.

Voorbeeld



Een doosje wordt aselect gekozen en uit de gekozen doos wordt aselect één knikker getrokken.
We bekijken de gebeurtenissen:
U: de getrokken knikker is wit
N: de getrokken knikker is zwart
Hoe groot is P(N)?
Uit het kansendiagram volgt

$$P(N) = \frac{1}{3}\cdot\frac{7}{10} + \frac{1}{3}\cdot\frac{1}{2} + \frac{1}{3}\cdot\frac{1}{5} = \frac{7}{15}.$$

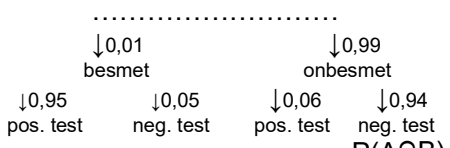


Stel eens dat de getrokken knikker zwart is. Hoe groot is dan de kans dat deze knikker uit doos V_1 komt?
Dit is een voorwaardelijke kans! N.L.

$$P(V_1/N) = \frac{P(V_1 \cap N)}{P(N)} = \frac{\frac{1}{3}\cdot\frac{7}{10}}{\frac{7}{15}} = \frac{1}{2}.$$

Een meer illustratief <u>voorbeeld</u> van het begrip voorwaardelijke kans.
Van een bepaald virus is bekend dat 1% van de bevolking ermee besmet is. Bij een bepaalde medische test reageert 95% van de besmetten positief en 6% van de onbesmetten.
Als iemand positief reageert, wat is dan de kans dat er sprake van besmetting is?

Kansendiagram:



$$P(\text{besmet} \mid \text{pos. test}) = P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0{,}01 \times 0{,}95}{0{,}01 \times 0{,}95 + 0{,}99 \times 0{,}06} = 0{,}138.$$

<u>Gevolg</u>
$P(A \cap B) = P(A)P(B|A)$
$P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B)$
Enz.

<u>Stelling</u>
$P(A|B) = P(A)$ d.e.s.d. als $P(A \cap B) = P(A) \times P(B)$.

Als dit het geval is, heten de gebeurtenissen A en B (stochastisch) <u>onafhankelijk</u>.

N.B.
Bij een enkelvoudig experiment kunnen, tegen de verwachting in, twee gebeurtenissen toch best onafhankelijk zijn!
Een simpel voorbeeld.
Neem een worp met een zuivere dobbelsteen.
$\Omega = \{1,2,3,4,5,6\}$
Beschouw de gebeurtenissen $A = \{2,3,5\}$, $B = \{2,4\}$ en $C = \{2,4,5,6\}$.
Dan geldt $P(A \cap B) = P(A) \times P(B)$ en $P(B \cap C) \neq P(B) \times P(C)$.

<u>Stelling</u>
Als $\Omega = \cup A_i$ met A_i (i=1 t/m n) onderling disjunct, dan $P(B) = \sum_i P(A_i)P(B|A_i)$.

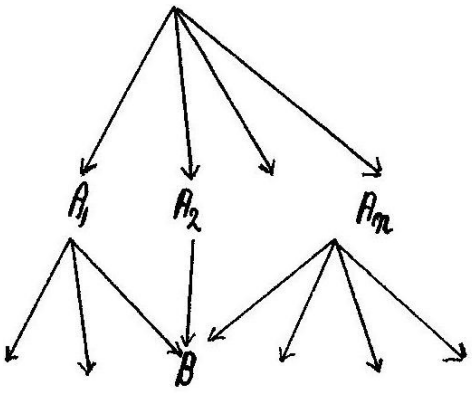
Bewijs:

$$B = \Omega \cap B = \left(\bigcup_{i=1}^{n} A_i\right) \cap B = \bigcup_{i=1}^{n} (A_i \cap B), \text{ dus volgens samregel}$$

($A_i \cap B$: disjunct)

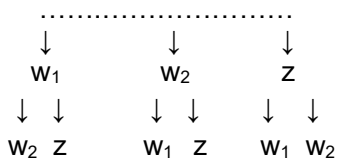
$$P(B) = \sum_{i=1}^{n} P(A_i \cap B) = \sum_{i=1}^{n} P(A_i) P(B|A_i)$$

Belang van deze stelling:
deze stelling rechtvaardigt het rekenen met een <u>kansendiagram</u>.

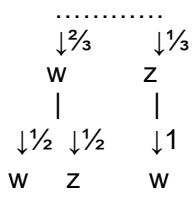


Om bij B te komen, moet men eerst een der A_i passeren. De kans om bij B te komen is dan uit te rekenen door voor elk van de gunstige paden naar B de kansen langs dat pad te vermenigvuldigen en dan de resultaten op te tellen.

Nog een <u>voorbeeld</u> om dit duidelijk te maken.
In een bakje zitten 2 witte balletjes (w_1, w_2) en 1 zwart balletje (z).
Trek eerst een balletje, zonder teruglegging.
Trek daarna nog een keer een balletje.
Boomdiagram:



Dit zijn 3 x 2 = 6 gelijkwaardige wegen.
Volgens de definitie van Laplace geldt:
P(ww) = 2/6, P(wz) = 2/6, P(zw) = 2/6
P (2 verschill. kl.) = 4/6 en P(2 gelijke kl.) = 2/6
Met een <u>kansendiagram</u> gaat dit echter sneller:



P(ww) = ⅔ x ½ = ⅓, P(wz) = ⅔ x ½ = ⅓, P(zw) = ⅓ x 1 = ⅓
P (2 verschill. kl.) = ⅔ x ½ + ⅓ x 1 = ⅔ en P(2 gelijke kl.) = ⅔ x ½ = ⅓

<u>Stelling van Bayes</u>

Als $\Omega = \bigcup_i A_i$ met A_i disjunct, dan $P(A_j/B) = \dfrac{P(A_j)P(B/A_j)}{\sum_i P(A_i)P(B/A_i)}$

want $P(A_j/B) = \dfrac{P(A_j \cap B)}{P(B)} = \dfrac{P(A_j)P(B/A_j)}{\sum_i P(A_i)P(B/A_i)}$

Een zeer belangrijke stelling!
De strekking is: aangenomen dat ik in B ben, wat is dan de kans dat ik over A_j ben gekomen.
Hierop geeft deze stelling antwoord.
Of nauwkeuriger: aangenomen dat sprake is van B, wat is dan de kans op A_j ?

Illustratief voorbeeld:
Een arts ziet bij een patiënt symptoom B. Dit symptoom kan het gevolg zijn van verschillende ziektes A_1, ….., A_n. De arts interesseert zich echter voor de kans op ziekte A_j, dus voor $P(A_j|B)$.

Twee toevalsexperimenten heten (fysisch) onafhankelijk als de uitkomsten van het ene experiment niet van invloed zijn op de uitkomsten van het andere experiment.
Gebeurtenissen bij het ene experiment en het andere experiment zijn dan (stochastisch) onafhankelijk van elkaar.

2. Combinatoriek

Stelling: somregel voor tellen
Als Ω op te splitsen is in disjuncte deelverzamelingen A_1 t/m A_r, dan geldt voor het aantal elementen van Ω:
$N(\Omega) = N(A_1 \cup A_2 \cup \ldots\ldots \cup A_r) = N(A_1) + N(A_2) + \ldots\ldots\ldots N(A_r)$

Voorbeeld
Bij twee worpen met een zuivere dobbelsteen zijn er 36 fysisch gelijkwaardige uittkomsten mogelijk.
Bij hoeveel uitkomsten is het verschil tussen beide worpen hoogstens 2 ogen?

Antwoord:
We verdelen de gunstige uitkomsten in drie porties:
- het verschil tussen beide worpen is 0;
  gunstige uitkomsten zijn dan 11, 22, 33, 44, 55 en 66. Dus 6 stuks.
- het verschil tussen beide worpen is 1;
  gunstig zijn dan 12 en 21, 23 en 32, ……, 56 en 56. Dus 10 stuks.
- het verschil tussen beide worpen is 2;
  gunstig zijn 13 en 31, 24 en 42, 35 en 53, 46 en 64. Dus 8 stuks.

Het antwoord is dus: 6 + 10 + 8 = 24.

Stelling: produktregel voor tellen
Als een meervoudig experiment uit r onafhankelijke deelexperimenten bestaat met resp. n_1, ……, n_r mogelijke uitkomsten, dan heeft het meervoudige experiment n_1 x …… x n_r mogelijke uitkomsten.

Stelling: permutaties
n verschillende objecten zijn op n! = n(n-1)(n-2)……..1 manieren op volgorde te zetten.

Stelling: binomiaalgetallen
Een verzameling met n elementen heeft $\binom{n}{s} = \frac{n(n-1)(n-2)\ldots\ldots\ldots\ldots(n-s+1)}{s(s-1)(s-2)\ldots\ldots\ldots 3.2.1}$ deelverzamelingen met s elementen.
Direct is in te zien dat $\binom{n}{s} = \frac{n!}{s!(n-s)!}$ en $\binom{n}{s} = \frac{n}{s}\binom{n-1}{s-1}$
Ook geldt $\binom{n}{s} = \binom{n-1}{s-1} + \binom{n-1}{s}$

Voorbeeld
In een vaas zitten 5 rode, 4 witte en 3 blauwe knikkers.
Er wordt vier keer met teruglegging getrokken.
Hoe groot is de kans dat daarbij 2 keer rood, 1 keer wit en 1 keer blauw wordt getrokken?

Antwoord:
In de vaas zitten 12 knikkers, die alle evenveel kans lopen om getrokken te worden. Het totale aantal gelijkwaardige uitkomsten is dus 12 x 12 x 12 x 12 = 12^4.
Hoeveel daarvan zijn gunstig?
Deze telling doen we in stappen.
1e stap: kies de 2 trekkingen die rood zullen opleveren. Dat zijn $\binom{4}{2}$ = 6 trekkingen.
2e stap: kies de 2 rode knikkers die daarbij getrokken zullen worden. Daarvoor zijn 5 x 5 = 25 mogelijkheden.

3e stap: kies uit de resterende 2 trekkingen 1 trekking voor wit. Beide trekkingen zijn dus mogelijk.
4e stap: kies welke witte knikker daarbij gekozen zal worden. Daarvoor zijn dus 4 mogelijkheden.
5e stap: kies de blauwe knikker voor de overblijvende trekking. Aantal mogelijkheden: 3.
In totaal zijn er dus 6 x 25 x 2 x 4 x 3 gunstige uitkomsten.
De gevraagde kans is dus 6 x 25 x 2 x 4 x 3 / 12^4 = 25 / 144.

Stelling: rangschikking

Men kan a nullen en b enen op $\binom{a+b}{a} = \binom{a+b}{b}$ manieren op volgorde zetten.

Voorbeeld

Hoe groot is de kans in 10 worpen met een zuivere munt precies 5 keer kruis te gooien?
Oplossing:
Mogelijke uitkomsten: alle rijtjes met 10 nullen / enen.
Gunstige uitkomsten: alle rijtjes met 5 nullen en 5 enen.

Dus gevraagde kans is $\binom{10}{5}$ / 2^{10} = 252 / 1024 = 0,246.

Stelling: verdeling

Men kan s gelijke knikkers op $\binom{n+s-1}{s}$ manieren over n verschillende vazen verdelen.
Immers, dit komt neer op s nullen tussen n – 1 tussenschotjes (enen) plaatsen.

Voorbeeld

Op hoeveel manieren kunnen 60 zetels over 3 partijen worden verdeeld?

Antwoord:

Dus s = 60 en n = 3, zodat het antwoord $\binom{62}{60} = \binom{62}{2}$ = 1891 is.
En nu de vraag: op hoeveel manieren kunnen 150 zetels over 18 partijen verdeeld worden? (Verkiezingsuitslag Nederland 2021)

Meer voorbeelden

Het driedeurenprobleem.

Een probleem dat publieke bekendheid kreeg door spelshows op tv, o.a. van Monty Hall in Amerika en Willem Ruis in Nederland.
Het probleem komt op het volgende neer:

De quizwinnaar mag kiezen uit drie deuren. Achter een van de drie deuren zit een grote prijs, achter de andere twee deuren zit niks. Als de deelnemer nu een van de drie deuren aanwijst (maar nog niet geopend heeft), opent de quizmaster (die weet waar de prijs zit) een deur waarachter de prijs niet zit. De quizmaster biedt de deelnemer de gelegenheid alsnog de overblijvende deur te kiezen.
Wat moet de deelnemer nu doen?

De meeste mensen denken intuïtief dat dit niets uitmaakt. De quizmaster opent een deur waar geen prijs achter zit. De prijs zit dus achter een van de twee andere deuren, waarvan een al aangewezen is door de deelnemer. Wat maakt het dan uit of je nog wisselt?
Ja ja, hierbij veronderstel je gemakshalve dat de kansen voor beide deuren gelijk zijn.
Hier laat de intuïtie je echter fors in de steek! Als je de moeite neemt om wat langer na te denken, dan gaat het volgende duidelijk worden:
De deelnemer kiest eerst voor een van de drie deuren. De kans op de prijs is dan 1/3.
Als je niet wisselt, blijft deze kans onveranderd.

Als je wel wisselt, dan heb je de prijs als je eerst een foute deur aangewezen hebt. De kans hierop was 2/3.
Dus, als je wel wisselt, dan neemt de kans op de prijs toe van 1/3 naar 2/3!

Dit is een prachtig voorbeeld hoe de intuïtie je kan bedriegen.
Voor veel mensen blijft het moeilijk om te geloven dat de intuïtieve redenering fout is, ook na de nodige uitleg. De uitlegger loopt dan het risico dat de tegenpartij denkt voor de gek gehouden te worden en dan kwaad wordt.
Op internet is zeer veel informatie over dit bijzondere probleem te vinden.

Het verjaardagenprobleem
In een zaal bevinden zich n personen. Hoe groot is de kans dat minstens twee personen op dezelfde dag jarig zijn?

Antwoord:
We nemen aan dat een jaar 365 dagen heeft, die als verjaardag even waarschijnlijk zijn.
De kans dat minstens twee personen op dezelfde dag jarig zijn is volgens de complementregel gelijk aan:
1 minus de kans dat alle n personen op verschillende dagen jarig zijn =

$$1 - \frac{365\ x\ 364\ x\ 363\ x \ldots\ldots x\ (365-n+1)}{365^n} = f(n)$$

Hier volgt een tabel met functiewaarden:

n	f(n)	n	f(n)	n	f(n)	n	f(n)
10	0.11695	28	0.65446	46	0.94825	64	0.99719
11	0.14114	29	0.68097	47	0.95477	65	0.99768
12	0.16702	30	0.70632	48	0.9606	66	0.9981
13	0.19441	31	0.73045	49	0.96578	67	0.99844
14	0.2231	32	0.75335	50	0.97037	68	0.99873
15	0.2529	33	0.77497	51	0.97443	69	0.99896
16	0.2836	34	0.79532	52	0.978	70	0.99916
17	0.31501	35	0.81438	53	0.98114	71	0.99932
18	0.34691	36	0.83218	54	0.98388	72	0.99945
19	0.37912	37	0.84873	55	0.98626	73	0.99956
20	0.41144	38	0.86407	56	0.98833	74	0.99965
21	0.44369	39	0.87822	57	0.99012	75	0.99972
22	0.4757	40	0.89123	58	0.99166	76	0.99978
23	0.5073	41	0.90315	59	0.99299	77	0.99982
24	0.53834	42	0.91403	60	0.99412	78	0.99986
25	0.5687	43	0.92392	61	0.99509	79	0.99989
26	0.59824	44	0.93289	62	0.99591	80	0.99991
27	0.62686	45	0.94098	63	0.9966		

Bij 23 personen is de kans al meer dan 50 % ! Bij 41 personen al meer dan 90 %.

Niet te verwarren met een ander verjaardagenprobleem:
In een zaal bevinden zich, buiten mij, nog n personen.
Hoe groot is de kans dat minstens één persoon tegelijk met mij jarig is?

Antwoord:

$$1 - \left(\frac{364}{365}\right)^n = g(n)$$

Enige functiewaarden:

n	g(n)
10	0,0271
23	0,0612
41	0,1064
253	0,5005

Dus g(n) << f(n)!

Het krentenprobleem

Dit probleem is heel mooi m.b.v. toevalscijfers te simuleren. Kies of maak een lijst met 100 paren toevalscijfers (x,y), waarbij x en y variëren van 0 t/m 9.
Maak een schema van 10 bij 10 velden. Elk van de 100 paren levert dan een kruisje in een van de 100 velden op.
Resultaat van deze simulatie:
Lijst van 100 paren toevalscijfers:

9	3	4	7	3	9	7	0	3	3
3	2	9	7	9	2	7	6	7	5
9	6	7	7	0	6	3	4	8	5
8	7	0	7	3	8	0	1	1	4
1	6	6	4	8	2	3	5	7	4
8	4	6	4	1	7	5	8	0	3
6	1	7	8	2	9	1	1	1	5
3	6	0	5	9	4	4	0	6	1
7	1	2	4	0	9	4	8	9	1
4	2	7	8	2	5	7	6	9	8
2	3	8	2	6	9	0	9	2	0
7	7	5	0	8	3	0	8	6	3
5	3	2	2	7	2	2	9	1	6
7	6	5	8	2	7	8	1	4	6
4	9	2	1	5	4	9	6	5	8
1	7	0	9	9	6	1	3	1	3
4	2	6	9	7	7	1	5	6	6
6	5	1	4	5	5	4	4	7	4
1	0	9	3	7	1	3	8	2	0
2	4	4	5	1	7	1	3	7	9

Invulling in de 100 velden levert als resultaat:

9	XXX		XX	X	X		XX	X		
8	X			XX	X	XXX		XX		X
7	X	XXX	X		X			XXX	X	X
6	X	XX		X	X		X	XXX		XXX
5	X	XX	X	X	X	X	X	X	X	
4		XX	XX	X	X	X	XX	XX	X	X
3	X	XXX	X	X		X	X		X	XX
2			X	X	XX			X	XX	X
1	X	X	X				XX	XX	X	X
0		X	XX		X	X		X		
	0	1	2	3	4	5	6	7	8	9

We zien dan:
30 velden met 0 kruisjes, 47 velden met 1 kruisje, 16 velden met 2 kruisjes en 7 velden met 3 kruisjes.
De verdeling van de krenten is dus nogal anders dan u wellicht zou verwachten.
De theoretische verdeling is ook eenvoudig te berekenen.
We moeten 100 keer een kruisje in een van de velden zetten. Dat betekent 100 plaatsingen na elkaar uitvoeren.
Neem een bepaald veld V in gedachten.
Als hier geen kruisje in mag, dan moeten alle 100 kruisjes in een van de andere 99 velden komen. Dat kan op 99^{100} manieren.
Als in veld V slechts 1 kruisje mag, dan moeten we eerst bepalen bij welk van de 100 plaatsingen dit gebeurt. Hier zijn 100 kandidaten voor. De andere 99 plaatsingen zijn bestemd voor de andere 99 velden.
Als in veld V 2 kruisjes mogen, dan moeten we nu bepalen bij welke 2 van de 100 plaatsingen dit zal gebeuren. Hier zijn 100 boven 2 kandidaten voor. De andere 98 plaatsingen zijn dan bestemd voor de 99 andere velden.
Enz.
Zo komen we tot de volgende kansen:

$$P(\text{0 kruisjes in veld V}) = \frac{99^{100}}{100^{100}} \approx 0{,}366$$

$$P(\text{1 kruisje in veld V}) = \frac{\binom{100}{1} \times 99^{99}}{100^{100}} \approx 0{,}370$$

$$P(\text{2 kruisjes in veld V}) = \frac{\binom{100}{2} \times 99^{98}}{100^{100}} \approx 0{,}185$$

$$P(\text{3 kruisjes in veld V}) = \frac{\binom{100}{3} \times 99^{97}}{100^{100}} \approx 0{,}061$$

$$P(>3 \text{ kruisjes in veld V}) = 1 - 0{,}366 - 0{,}370 - 0{,}185 - 0{,}061 \approx 0{,}018$$

Dus de theoretische verdeling wordt:

aantal velden met 0 kruisjes:	36,6
aantal velden met 1 kruisje:	37
aantal velden met 2 kruisjes:	18,5
aantal velden met 3 kruisjes:	6,1
aantal velden met >3 kruisjes:	1,8

Het kan zeer verhelderend zijn als u zelf ook eens deze simulatie uitvoert.
Toevalsexperimenten simuleren m.b.v. toevalscijfers wordt vaak de Monte-Carlo-methode genoemd.

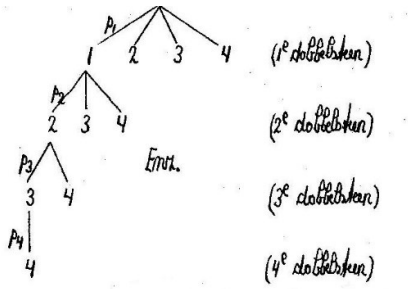
3. Interessante problemen uit de geschiedenis

Dobbelspel uit de klassieke oudheid

De kansrekening is ontstaan uit gokspelen. Kansspelen bloeiden ten allen tijde en in veel culturen. Reeds in het stenen tijdperk was dobbelen wijd verbreid. Men zou dus niet verwachten dat deze tak van wetenschap pas zo'n 300 jaar geleden tot ontwikkeling begon te komen.
Hieronder een vaas uit omstreeks 600 v. Chr., met Achilles en Ajax aan het dobbelen.

Als dobbelstenen gebruikte men bepaalde dierbeenderen (gewrichten), astragalia genaamd. Zulke astragalia kunnen op 4 zijden vallen.
Afbeelding:

Experimenteren met museumexemplaren leverde kansen op zoals $p_1 \approx p_2 \approx 0{,}4$ en $p_3 \approx p_4 \approx 0{,}1$.
Meestal werd met vier astragalia tegelijk geworpen. Een worp met vier verschillende zijden boven heette Venusworp. Dat was iets bijzonders.
Hoe groot is nu de kans op een Venusworp?
We kunnen deze kans uitrekenen met een gereduceerd kansendiagram:



Het pad 1 2 3 4 heeft kans $p_1 \times p_2 \times p_3 \times p_4$.
Er zijn 4! van deze paden. Dus de gevraagde kans is $24 \times 0{,}4 \times 0{,}4 \times 0{,}1 \times 0{,}1 \approx 0{,}04$.

Het probleem van Galilei (1564 – 1642)
De vorst van Toscane vroeg aan Galilei: waarom verschijnt bij een worp met drie dobbelstenen de ogensom 10 vaker dan de ogensom 9, hoewel beide ogensommen op zes manieren kunnen intreden?
Immers, 9 = 1+2+6 = 1+3+5 = 1+4+4 = 2+2+5 = 2+3+4 = 3+3+3 en
10 = 1+3+6 = 1+4+5 = 2+2+6 = 2+3+5 = 2+4+4 = 3+3+4.
Dit was destijds een veelbesproken eeuwenoud probleem.
Galilei vond de verklaring: de zes manieren zijn niet even waarschijnlijk!
Het spel heeft $6^3 = 216$ mogelijke uitkomsten, die even waarschijnlijk zijn.
Voor ogensom 9 zijn 25 uitkomsten gunstig, voor ogensom 10 daarentegen 27.
Dus hebben ogensom 9 en 10 resp. kans $25/216 \approx 0{,}116$ en $27/216 \approx 0{,}125$.
Het verschil is klein en kan experimenteel maar moeilijk worden vastgesteld.

De problemen van Chevalier de Méré (1607 – 1684)
Franse ridder, schrijver, gokker en filosoof (aan het hof van Lodewijk de 14e).
Legde in 1654 twee problemen voor aan de bekende wiskundige Blaise Pascal (1623 – 1663).

1) Wat is waarschijnlijker, minstens één zes bij 4 worpen met 1 dobbelsteen of minstens een dubbele zes bij 24 worpen met 2 dobbelstenen?
2) Een munt wordt herhaaldelijk geworpen. Bij kruis krijgt A 1 punt, bij munt krijgt B 1 punt. Wie het eerst 5 punten heeft, krijgt de inzet.
   Na 7 worpen heeft A 4 punten en B 3 punten. Het spel wordt afgebroken. Hoe moet de inzet worden verdeeld?

Discussies over dit en soortgelijke problemen hebben substantiëel bijgedragen aan het ontstaan van de kansrekening.

Deze twee vragen hebben tot een briefwisseling tussen Pascal en Fermat (1601 - 1665) geleid.

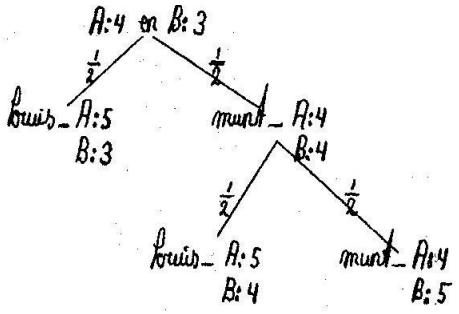
Probleem 1 is eenvoudig.
4 worpen met 1 dobbelsteen: P(minstens 1 zes) = 1 – P(geen zes) = $1 - (5/6)^4 \approx 0{,}5177$.
24 worpen met 2 dobbelstenen: P(minstens 1 dubbele zes) = 1 – P(geen dubbele zes) = $1 - (35/36)^{24} \approx 0{,}4914$.
Een miniem verschil dus! Beide kansen zijn praktisch gelijk aan 1/2.

Probleem 2 is lastiger.
Twee vaak gehoorde oplossingen waren:

- A en B verdelen de inzet in de verhoudig 4 : 3, in overeenstemming met de reeds behaalde punten.
- A en B verdelen de inzet in de verhouding 2 : 1, in overeenstemming met de nog ontbrekende punten.

Pascal redeneerde echter als volgt:

Stel dat de inzet 64 pistolen was. Met kans 1/2 wint A deze inzet bij de volgende worp. Dus komen hem alvast 32 pistolen toe. Als A daarentegen de volgende worp verliest, dan hebben beiden evenveel (4) punten en dus daarna evenveel kans om te winnen, dus van de overige 32 pistolen komen A 16 pistolen toe. De juiste verdeling van de inzet is dus 48 pistolen voor A en 16 voor B. Dat is 3 : 1.
Een kansendiagram laat ook zien dat deze verdeling wiskundig voor de hand ligt:



De winkans voor A is 1/2 + 1/2 x 1/2 = 3/4 en voor B 1/2 x 1/2 = 1/4. De verdeling van de inzet in de verhouding 3 : 1 is dus redelijk.

<u>De worsteling met het begrip kans nog na 1700</u>
Zelfs in de simpelste kansprocessen zaten toen nog valkuilen waar zelfs grote wiskundigen nog in trapten. De beroemde Duitse filosoof en wiskundige Leibniz (G.W. Leibniz, 1646-1716) schreef in 1715 dat je met twee dobbelstenen net zo makkelijk 11 als 12 gooit, want beide uitkomsten zouden maar op een manier te gooien zijn: respectievelijk als 5;6 en als 6;6.

4. **Toevalsvariabelen**

Fucties op de uitkomstenverzameling Ω heten toevalsvariabelen. Een toevalsvariabele is een functie X die aan elke mogelijke uitkomst ω een getal X(ω) toekent. Bij elke mogelijke waarde x (x is een gebeurtenis) hoort een kans p(x). De tabel x / p(x) heet de kansverdeling van de toevalsvariabele.
De verwachtingswaarde wordt gedefiniëerd als $E(X) = \sum X(\omega)p(\omega) = \sum xp(x)$.

We beschouwen het voorbeeld van 2 worpen met een zuivere dobbelsteen. De uitkomstenverzameling Ω = {11,12,13,……,66} bestaat uit $6^2 = 36$ elementen. De totale ogensom van de 2 worpen is een toevalsvariabele. De mogelijke waarden zijn dus x = 2, 3, …., 12. De kansverdeling is:

X	2	3	4	5	6	7	8	9	10	11	12
p(x)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Dus $E(X) = \sum xp(x) = 2 \times 1/36 + 3 \times 2/36 + \ldots\ldots. + 12 \times 1/36 = 7$.

Voorbeeld
A bezit een villa ter waarde van € 500.000. De jaarlijkse brandverzekeringspremie is € 400. De kans op brand voor dit type huizen is, volgens statistieken van de assurantiemaatschappij, ca. 0,05 % (per jaar gerekend).
Zij X het jaarlijks verlies zonder brandverzekering en Y het jaarlijks verlies met brandverzekering.
Dan geldt E(X) = 500.000 x 0,0005 = 250 en E(Y) = 400.
Verzekeren is dus een ongunstig spel. Toch kiest A terecht voor verzekeren, omdat het risico anders klein, maar ondraaglijk is.

Voorbeeld: groepsgewijs onderzoek
Stel dat we een zeer groot aantal personen moeten onderzoek op het al dan niet hebben van een bepaalde aandoening en dat deze aandoening door middel van bloedonderzoek kan worden aangetroond.
Hoe kunnen we dit aanpakken?
Laat q het aandeel van de bevolking zijn dat deze aandoening heeft. Dan is p = 1 – q de fractie van de bevolking zonder deze aandoening.
We vergelijken twee methoden van onderzoek.

1) Individueel onderzoek

Elke persoon wordt afzonderlijk onderzocht. Per persoon is dan 1 bloedonderzoek nodig.

2) Groepsgewijs onderzoek

Het bloed van r personen wordt gemengd en onderzocht. Met kans p^r zijn allen gezond en is slechts 1 bloedonderzoek nodig. Met kans $1 – p^r$ heeft minstens één persoon de aandoening en in dat geval moet iedere persoon uit deze groep alsnog afzonderlijk worden onderzocht. In dat geval zijn nog r bloedonderzoeken nodig.

Per groep van r personen moeten we dan verwachten dat $1 \times p^r + (1+r) \times (1 – p^r) =$ $1 + r(1 – p^r)$ bloedonderzoeken nodig zullen zijn. Dat is $1/r + 1 – p^r$ onderzoeken per persoon.

Bij groepsgewijs onderzoek is dus een besparing te verwachten van $p^r – 1/r$ onderzoeken per persoon.
Wat is bij gegeven p nu het maximum van deze besparing?
Een tabel geeft:

p	0,7	0,8	0,85	0,90	0,91	0,92	0,93	0,94	0,95	0,96	0,97	0,98	0,99
r optimaal	3	3	3	4	4	4	5	5	5	6	6	8	11
besparing (in %)	1	18	28	41	44	47	50	53	57	62	67	73	80

N.B.
Tijdens de tweede wereldoorlog moesten in de VS miljoenen recruten in korte tijd medisch onderzocht worden, o.a. op syphilis. Rond 1 % van de mannen in deze leeftijdscategorie had syphilis, dus $p \approx 0{,}99$. Bovenstaande tabel laat zien dat het onderzoek het beste in groepen van 11 kon plaatsvinden. De daarbij bereikte besparing was dan 80 %.

Stelling
Als X en Y toevalsvariabelen op dezelfde uitkomstenverzameling zijn en a en b reële getallen, dan geldt $E(aX+bY) = aE(X) + bE(Y)$.

Bewijs:
$E(aX+bY) = \sum\{aX(\omega) + bY(\omega)\}p(\omega) = a\sum X(\omega)p(\omega) + b\sum Y(\omega)p(\omega) = aE(X) + bE(Y)$.

Er geldt niet algemeen $E(XY) = E(X)E(Y)$!

Twee toevalsvariabelen X en Y op dezelfde uitkomstenverzameling heten onafhankelijk als $P(X=x \wedge Y=y) = P(X=x)P(Y=y)$ voor alle mogelijke waarden van x en y.

Stelling
Als X en Y onafhankelijke toevalsvariabelen op dezelfde uitkomstenverzameling zijn, dan geldt $E(XY) = E(X)E(Y)$.

Bewijs:
Uitschrijven voor een eenvoudig voorbeeld van twee onafhankelijke toevalsvariabelen is direct overtuigend.

Deze stelling is zeer handig.
Voorbeeld
Experiment: 2 worpen met een zuivere dobbelsteen.
Toevalsvariabelen:
X: ogensom 1^e worp
Y: ogensom 2^e worp
De mogelijke waarden van X en Y zijn dan voor beide 1,2,3,4,5,6.
De toevalsvariabelen X en Y zijn onafhankelijk. Bijv. $P(X=3 \wedge Y=5) = 1/36 = 1/6 \times 1/6 = P(X=3).P(Y=5)$. Enz.

Dus we mogen de stelling toepassen:

$E(XY) = E(X)E(Y) = \left(1.\frac{1}{6} + 2.\frac{1}{6} + 3.\frac{1}{6} + 4.\frac{1}{6} + 5.\frac{1}{6} + 6.\frac{1}{6}\right)^2 = \left(\frac{7}{2}\right)^2 = 12\frac{1}{4}$.

Konden we niet over deze stelling beschikken, dan verliep de berekening aanzienlijk gecompliceerder:

$E(XY) = 1.\frac{1}{36} + 2.\frac{2}{36} + 3.\frac{2}{36} + 4.\frac{3}{36} + 5.\frac{2}{36} + 6.\frac{4}{36} + 8.\frac{2}{36} + 9.\frac{1}{36} + 10.\frac{2}{36} + 12.\frac{4}{36} + 15.\frac{2}{36} + 16.\frac{1}{36} + 18.\frac{2}{36} +$

$20.\frac{2}{36} + 24.\frac{2}{36} + 25.\frac{1}{36} + 30.\frac{2}{36} + 36.\frac{1}{36} = \frac{441}{36} = 12\frac{1}{4}$

gunstige uitkomsten voor waarde 12:
26, 34, 43 en 62

We kunnen ook onafhankelijkheid van meer dan twee toevalsvariabelen definiëren.

Drie toevalsvariabelen X, Y en Z op dezelfde uitkomstenverzameling heten <u>onafhankelijk</u> als $P(X=x \wedge Y=y \wedge Z=z) = P(X=x)P(Y=y)P(Z=z)$ voor alle mogelijke waarden van x, y en z.

Analoog voor meer dan drie toevalsvariabelen.

Zij $A \subset \Omega$.
De <u>indicatorfunctie</u> van A is de functie die aan elk element van A de waarde 1 toekent en aan elk element buiten A de waarde 0. Notatie: I_A.
Dus $I_A(\omega) = 1$ als $\omega \in A$
$= 0$ als $\omega \notin A$
Dan volgt:

$E(I_A) = P(A)$…….. (1)

Indicatorfuncties kunnen het berekenen van verwachtingswaarden gemakkelijker maken.
<u>Voorbeeld</u>
n gelijke knikkers worden aselect over 3 vazen verdeeld. Zij X het aantal lege vazen.
Bereken E(X).
De kansverdeling ziet er niet direct eenvoudig uit:
x = 0, 1, 2
p(x) = …… ?
Handig is dan X als som van indicatorfuncties te schrijven. Wel, als A, B en C de gebeurtenissen zijn dat resp. de eerste, tweede en derde vaas leeg blijven, dan geldt
$X = I_A + I_B + I_C$
Ga dit zelf eens na voor bijv. n = 3. Er geldt:

	waarde van X	waarden van de indicatorfuncties
geen vaas leeg	0	alle 0
1 vaas leeg	1	precies één heeft waarde 1
2 vazen leeg	2	twee hebben waarde 1

Dus $E(X) = E(I_A + I_B + I_C) = E(I_A) + E(I_B) + E(I_C) = P(A) + P(B) + P(C) = (2/3)^n + (2/3)^n + (2/3)^n =$
(E lineair)
$3(2/3)^n$. Klaar!

Stelling

$I_A^2 = I_A$	(2)
$I_{\overline{A}} = 1 - I_A$	(3)
$I_{A_1 \cap A_2 \cap \ldots \cap A_m} = I_{A_1} \cdot I_{A_2} \ldots I_{A_m}$	(4)
$I_{A_1 \cup A_2 \cup \ldots \cup A_m} = 1 - (1 - I_{A_1})(1 - I_{A_2}) \ldots (1 - I_{A_m})$	(5)

Bewijs van eig. 2:
als $\omega \in A$, dan $I_A(\omega) = 1 = 1^2 = I_A(\omega)^2 = I_A^2(\omega)$; als $\omega \notin A$, dan $I_A(\omega) = 0 = 0^2 = I_A(\omega)^2 = I_A^2(\omega)$.
Bewijs van eig. 3:
$1 - I_A(\omega)$ is precies dan 1 als $\omega \notin A$, en anders 0.
Bewijs van eig. 4:
het rechterlid is precies dan 1 als alle factoren I_{A_i} 1 zijn, d.w.z. als ω in elke verzameling A_i ligt, en anders 0.
Bewijs van eig. 5:
het rechterlid is precies dan 1 als het haakjesproduct 0 is, dus als minstens één der factoren $1 - I_{A_i}$ 0 is; dit is het geval als minstens één der verzamelingen A_i ω bevat. Zo niet, dan is het rechterlid 0.

Zij X een toevalsvariabele met:
mogelijke functiewaarden x_i (i=1 t/m n) en bijbehorende kansen $p(x_i)$.
De verwachtingswaarde $\mu = E(X) = \sum x_i p(x_i)$ is het gewogen gemiddelde van de functiewaarden, maar zegt nog niets over de verdeling van de functiewaarden. Daarvoor hebben we nog een andere maat nodig. Daarvoor wordt de verwachtingswaarde van $(X-\mu)^2$ genomen. Dit getal wordt de variantie van X genoemd. Notatie: V(X).
We hebben dus de volgende definitie:

$$V(X) = E\big((X-\mu)^2\big) = \sum (x_i-\mu)^2 p(x_i)$$

Omdat $(X-\mu)^2 = X^2 - 2\mu X + \mu^2$, geldt $V(X) = E(X^2) - 2\mu E(X) + E(\mu^2) = E(X^2) - 2\mu^2 + \mu^2$, dus

$$V(X) = E(X^2) - \mu^2 = E(X^2) - E(X)^2.$$

Onder de spreiding of standaarddeviatie wordt nu verstaan: $\sigma = \sqrt{V(X)}$.

Voorbeeld
Zij X het aantal keren munt bij een worp met 3 zuivere munten.
Dan is de kansverdeling van X de volgende:

x	0	1	2	3
p(x):	1/8	3/8	3/8	1/8

Dus E(X) = 0x1/8 + 1x3/8 + 2x3/8 + 3x1/8 = 1½

$E(X^2) = 0^2 \times 1/8 + 1^2 \times 3/8 + 2^2 \times 3/8 + 3^2 \times 1/8 = 3$, $V(X) = E(X^2) - \mu^2 = 3 - (1\frac{1}{2})^2 = \frac{3}{4}$, zodat $\sigma = \frac{1}{2}\sqrt{3} = 0{,}87$.

Stelling
$V(a) = 0$
$V(X+a) = V(X)$
$V(aX) = a^2 V(X)$

Bewijs.
$V(a) = E(a^2) - E(a)^2 = a^2 \cdot 1 - (a \cdot 1)^2 = 0.$
$V(X+a) = E((X+a)^2) - E(X+a)^2 = E(X^2 + 2aX + a^2) - (E(X) + E(a))^2 = E(X^2) + 2aE(X) + a^2 - E(X)^2 - 2aE(X) - a^2 =$
$E(X^2) - E(X)^2 = V(X).$
$V(aX) = E(a^2X^2) - E(aX)^2 = a^2 E(X^2) - (aE(X))^2 = a^2 \{E(X^2) - E(X)^2\} = a^2 V(X).$

Stelling
Voor onafhankelijke toevalsvariabelen X en Y geldt: $V(X+Y) = V(X) + V(Y)$.

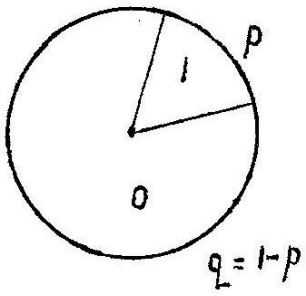
Bewijs.
$V(X+Y) = E(X^2 + 2XY + Y^2) - E(X+Y)^2$
$= E(X^2) + 2E(XY) + E(Y^2) - \{E(X) + E(Y)\}^2$ omdat E een lineaire operator is
$= E(X^2) + 2E(X)E(Y) + E(Y^2) - E(X)^2 - 2E(X)E(Y) - E(Y)^2$ omdat X en Y onafhankelijk zijn
$= V(X) + V(Y).$

Analoog voor elk eindig aantal onafhankelijke toevalsvariabelen.

Stelling
$V(I_A) = P(A)P(\neg A)$

Bewijs: is gemakkelijk na te gaan.

Nog een belangrijk voorbeeld.
Experiment: n herhalingen van hetzelfde experiment waarbij per keer de succeskans p is.
S_n: het aantal successen na deze n herhalingen.



Gevraagd: $E(S_n)$, $V(S_n)$ en $\sigma(S_n)$

Oplossing:
Elke uitkomst ω is een rij van nullen (missers) en enen (successen) ter lengte van n.
Als A_i de gebeurtenis "de i^e keer een 1" is en X_i is de indicatorfunctie van A_i, dan geldt
$S_n = X_1 + \ldots + X_n$ en dus
$E(S_n) = E(X_1 + \ldots + X_n) = E(X_1) + \ldots + E(X_n) = P(A_1) + \ldots + P(A_n) = np.$

De toevalsvariabelen X_1 t/m X_n zijn duidelijk onafhankelijk, dus $V(S_n) = V(X_1 + ... + X_n) = V(X_1) + ... + V(X_n) = P(A_1)P(\neg A_1) + ... + P(A_n)P(\neg A_n) = pq + ... + pq = npq$
Recapitulerend:

$E(S_n) = np$
$V(S_n) = npq$
$\sigma(S_n) = \sqrt{(npq)}$

Voorbeeld
Gegeven zijn 5 vazen. Vaas nr. i bevat 1 witte en i zwarte ballen. Uit elke vaas wordt aselect een knikker getrokken. Zij X het resulterende aantal witte knikkers. Bereken E(X) en σ(X).

Antwoord:

A_i is de de gebeurtenis dat de knikker uit de i^e vaas wit is.

Dan $X = I_{A_1} + \cdots + I_{A_n}$

$E(X) = E(I_{A_1} + \cdots) = E(I_{A_1}) + \cdots = P(A_1) + \cdots + P(A_5)$

$= \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} = 1{,}45$

$V(X) = V(I_{A_1} + \cdots) = V(I_{A_1}) + \cdots = P(A_1)P(\neg A_1) + \cdots$

$= \frac{1}{2} \times \frac{1}{2} + \frac{1}{3} \times \frac{2}{3} + \frac{1}{4} \times \frac{3}{4} + \frac{1}{5} \times \frac{4}{5} + \frac{1}{6} \times \frac{5}{6} = 0{,}9586$

dus $\sigma = \sqrt{0{,}9586} = 0{,}979$

Voorbeeld
Op de begane grond van een flat met n verdiepingen stappen s personen in de lift.
Zij X het aantal keren dat de lift moet stoppen. Bereken E(X) en V(X).

Antwoord:

A_i is de gebeurtenis dat de lift op de i^e verdieping moet stoppen.
Dan geldt:

$X = I_{A_1} + \cdots + I_{A_n}$

$P(A_i) = P(\text{minstens één persoon moet op de } i^e \text{ verdieping zijn})$
$= 1 - P(\text{niemand moet op de } i^e \text{ verdieping zijn}) = 1 - \frac{(n-1)^s}{n^s}$

$E(X) = E(I_{A_1} + \cdots) \underset{E \text{ lineair}}{=} E(I_{A_1}) + \cdots = P(A_1) + \cdots + P(A_n) = n\left\{1 - \left(\frac{n-1}{n}\right)^s\right\}$

$V(X) = V(I_{A_1} + \cdots) \underset{I_{A_i} \text{ onafh.}}{=} V(I_{A_1}) + \cdots = P(A_1)P(\neg A_1) + \cdots = \text{enz.}$

Deze voorbeelden maken duidelijk hoe handig het begrip indicatorfunctie is.

5. Toetsen van hypothesen

Voorbeeld 1
Kan een pasgeboren kuiken ronde graankorrels herkennen of leert het dit pas na de nodige ervaring?

Om deze vraag te beantwoorden wordt het volgende experiment opgezet. Zodra het kuiken uit het ei is, worden hem graankorrels voorgezet. De helft van de korrels is rond, de andere korrels zijn puntvormig. Heeft het pasgeboren kuiken voorkeur voor ronde korrels?
We stellen twee hypothesen op.
H_0 (nulhypothese): het kuiken heeft geen voorkeur voor ronde korrels. De kans dat een opgepikte korrel rond is is $p = ½$.
H_1 (tegenhypothese): het kuiken heeft voorkeur voor ronde korrels. De kans dat een opgepikte korrel rond is is $p > ½$.
Let wel: de mogelijkheid $p < ½$ komt niet in aanmerking, omdat deze niet plausibel is d.w.z. er is geen reden om hiermee rekening te houden.
We laten het kuiken 10 keer pikken.
Resultaat: 0 0 0 Δ 0 0 0 Δ 0 0
Wat mogen we nu concluderen?
Stel dat H_0 waar is. Dan heeft het kuiken dus geen voorkeur voor ronde korrels. Het kuiken heeft echter 8 van de 10 keer een ronde korrel gepikt. Is er dan toch sprake van voorkeur?
Laat X het aantal ronde korrels zijn na 10 keer pikken. Hoe groot moet X zijn om H_0 te kunnen verwerpen?

$(P(X \geq 9) = \frac{\binom{10}{9}+\binom{10}{10}}{2^{10}} \approx 0{,}011$ en $P(\geq 8) = \frac{\binom{10}{8}+\binom{10}{9}+\binom{10}{10}}{2^{10}} \approx 0{,}055$

Beslissingsregel: deze kans moet kleiner dan 5% zijn, willen we H_0 verwerpen.
De kritieke zone is: K = {9,10}.
Onbetrouwbaarheid of significantieniveau: $\alpha = P(X \geq 9) = 0{,}011$
Betrouwbaarheid of statistische zekerheid: $1-\alpha$

Als de waargenomen waarde X in de kritieke zone valt, dan spreken we van een significante afwijking.
Als het om een zwaarwegende beslissing gaat, dan kunnen we eventueel α kleiner nemen.

In bovenstaand voorbeeld spreken we van een rechts-eenzijdige toets.

Voorbeeld 2
Zijn ratten kleurenblind?
Om dit te onderzoeken wordt het volgende experiment bedacht:
10 ratten worden door een gang gestuurd, die zich in twee gangen splitst, een rood en een groen geverfde gang.
We stellen twee hypothesen op.
H_0 (nulhypothese): ratten hebben geen voorkeur voor een van beide kleuren (misschien omdat ze kleurenblind zijn); de kans dat ze de groene gang kiezen is $p = ½$.
H_1 (tegenhypothese): ratten hebben voorkeur voor een van beide kleuren (in dat geval zijn ze zeker niet kleurenblind), d.w.z $p \neq ½$. Dus $p > ½$ of $p < ½$.
Resultaat van het experiment: 8 ratten kiezen de groene gang.
Wat mogen we nu concluderen?

Laten we eens aannemen dat H_0 waar is. Als X het aantal ratten is dat groen kiest, dan pleiten grote waarden van X evenzeer tegen H_0 als kleine waarden van X.

P(X=0,1,2 of 8,9,10) = $\frac{\binom{10}{0}+\binom{10}{1}+\binom{10}{2}+\binom{10}{8}+\binom{10}{9}+\binom{10}{10}}{2^{10}} \approx$ 112/1024 = 0,109 (dus meer dan 10%).
P(X=0,1 of 9,10) = 22/1024 = 0,021 (dus minder dan 10%)
De kritieke zone is dus K = {0,1,9,10} en daar valt 8 dus nog net niet in.
Er is dus nog net geen reden om H_0 te verwerpen.

In dit voorbeeld spreken we van een tweezijdige toets en dan is het gebruikelijk H_0 pas te verwerpen als $\alpha < 0{,}10$.

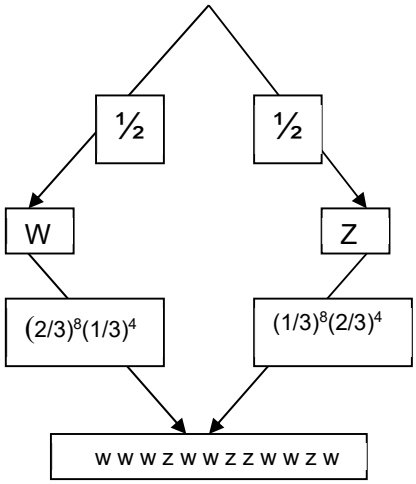
6. **Het statistische alternatief probleem**

Stel dat van de onbekende kans p slechts twee waarden p_1 en p_2 in aanmerking komen.

Voorbeeld

We hebben een vaas W = [o ☻ o] met 2 witte en 1 zwarte bal en een vaas Z = [☻ o ☻] met 2 zwarte en 1 witte bal.
We kiezen aselect een van beide vazen en trekken dan 12 keer een balletje met teruglegging.
Resultaat: o o o ☻ o o ☻ ☻ o o ☻ o
Welke van beide vazen was het nu? Er zijn duidelijk meer witte knikkers dan zwarte, dus dit resultaat pleit voor vaas W.

We onderscheiden de volgende hypothesen:
H_1: vaas W is gekozen
H_2: vaas Z is gekozen
Het kansendiagram wordt:



Deze wegen hebben resp. de kansen ½ $(2/3)^8(1/3)^4$ en ½ $(1/3)^8(2/3)^4$, d.i. $2^8 : 2^4 = 16 : 1$.
Dus H_1 en H_2 kunnen we resp. de kansen 16/17 en 1/17 toekennen.
Conclusie:
Als we beslist een keuze moeten maken tussen beide hypothesen, dan kiezen we natuurlijk voor de hypothese met de grootste kans, dus voor H_1.
Als deze noodzaak nier aanwezig is, dan zullen we in he algemeen pas voor een van beide hypothesen kiezen als de bijbehorende kans meer dan 95% is.

In dit voorbeeld kan men met willekeurig kleine onbetrouwbaarheid het type van de getrokken vaas bepalen. Men hoeft alleen maar voldoende veel trekkingen met teruglegging te verrichten. In de praktijk gaat dat echter vaak niet. Trekkingen kunnen bijv. erg duur zijn. Denk maar eens aan kwaliteitscontrole in de industrie. Er moet arbeid verricht worden om de controle uit te voeren en vaak is het nodig het produkt daarbij te vernielen. Men zal dus proberen met een minimaal aantal controles toe te kunnen.

We behandelen nu een voorbeeld waarbij tussen twee beslissingen gekozen moet worden met een minimaal aantal trekkingen. De hierbij gebruikte oplossingsmethode is afkomstig van de amerikaanse statisticus A. Wald (1902 – 1950).

Voorbeeld

Van een zeker geneesmiddel zijn twee kwaliteiten verkrijgbaar, kwaliteit A die in 75% van de gevallen genezing brengt, en kwaliteit B die in slechts 25% van de gevallen geneest.
Een laboratorium krijgt van een medisch centrum een doos ampullen van dit geneesmiddel toegestuurd met het verzoek vast te stellen om welke kwaliteit het hier gaat.
Een manier om dit te onderzoeken is een aantal witte muizen te besmetten en dan het geneesmiddel in te spuiten.
Het ligt voor de hand niet onnodig veel witte muizen in te spuiten. Daarbij komt nog dat de ampullen vrij duur zijn, dus men wil er zo weinig mogelijk van verspillen, zeker als het om kwaliteit A gaat.
Hoe gaat men te werk?
Wel, er zijn twee hypothesen in het geding:

H_1: de ampullen zijn van kwaliteit A
H_2: de ampullen zijn van kwaliteit B

Het geneesmiddel wordt ingespoten bij een aantal besmette witte muizen. We laten nog even in het midden hoeveel. Stel dat x muizen genezen en y muizen niet. Als x+y besmette muizen ingespoten zijn, dan is de kans op dit resultaat onder H_1 gelijk aan

$\binom{x+y}{x}(3/4)^x(1/4)^y$ en onder H_2 gelijk aan $\binom{x+y}{x}(1/4)^x(3/4)^y$. Deze kansen verhouden zich

als $(3/4)^x(1/4)^y : (1/4)^x(3/4)^y = 3^x : 3^y$.
Van te voren is beslist dat zodra voor één van beide hypothesen een kans van 99% pleit, gestopt wordt met het experiment en gekozen wordt voor de hypothese die een overtuigend resultaat geeft. Dus we gaan door zolang

$0{,}05 < P(H_1) = \frac{3^x}{3^x+3^y} < 0{,}95$ ofwel $|y-x| < 2{,}7$.

Dus zodra $|y-x| \geq 3$ kunnen we het experiment stoppen.

7. **Exacte test van Fisher**

R.A. Fisher, 1890 - 1962

Voorbeeld
Bij een bepaalde ziekte is een nieuw middel gevonden dat in een aantal gevallen werkzaam is. Een ziekenhuis gaat 10 patiënten met dit nieuwe middel behandelen en, ter vergelijking, 9 patiënten met een placebo (een middel zonder werkzame stof).
Resultaat:

	genezen	niet genezen
Nieuw middel	8	2
Placebo	5	4

Gebruikelijke terminologie in deze:
proefgroep, controlegroep (placebo-groep) en viervakkentabel.

De genezing is 8/10 = 0,80, tegen 5/9 = 0,56 in de placebo-groep.
De indruk is dus dat het nieuwe middel beter werkt dan een placebo. Mogen we dit inderdaad concluderen?
We stellen twee hypothesen op:
H_0: het nieuwe middel werkt niet, d.w.z. niet beter dan een placebo.
H_1: het nieuwe middel werkt wel, d.w.z. beter dan een placebo.
Stel H_0 is waar. Dan zullen dus 13 van de 19 patiënten genezen, ongeacht welk van de twee middelen ze krijgen. Onder de 10 patiënten die met het nieuwe middel behandeld zijn bevinden zich toevallig 8 van deze 13 patiënten. De kans dat in deze steekproef van 10 patiënten 8 of meer van deze 13 patiënten zitten is gelijk aan:

$$\frac{\binom{13}{8}\binom{6}{2}+\binom{13}{9}\binom{6}{1}+\binom{13}{10}\binom{6}{0}}{\binom{19}{10}} \approx 0{,}258$$

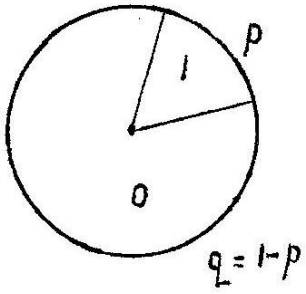
Dus $\alpha = 0{,}258$. D.w.z. de kans op een resultaat minstens zo opmerkelijk (pleitend voor H_1) als het waargenomen resultaat is ruim 25%. Dus er is geen reden om H_0 te verwerpen.

Voor het toetsen van hypothesen geldt algemeen:
Een fout van de 1^e soort: H_0 is waar en wordt toch verworpen. De kans om deze fout te maken is α.
Een fout van de 2^e soort: H_0 is onwaar en wordt toch aangenomen.

8. **Binomiale verdeling**

Beschouw het al eerder bekeken belangrijke voorbeeld:

Experiment: n herhalingen van hetzelfde experiment waarbij per keer de succeskans p is.
S_n: het aantal successen na deze n herhalingen.
Elke uitkomst ω is een rij van enen (successen) en nullen (missers).
Dan geldt: $E(S_n) = np$, $V(S_n) = npq$ en $\sigma(S_n) = \sqrt{(npq)}$.



Hoe ziet de kansverdeling van de toevalsvariabele S_n er nu uit?
S_n kan de waarden 0, 1, 2, …… , n aannemen. Wat zijn nu de kansen op deze waarden?

Als ω een rij van x enen en n-x nullen is, dan zijn $\binom{n}{x}$ van dergelijke rijen mogelijk. Elk van die rijen heeft kans $p^x q^{n-x}$, dus:

Stelling

$P(S_n = x) = b(x) = \binom{n}{x} p^x q^{n-x}$ met x = 0, 1, ….., n.

Dit heet de binomiale kansverdeling.

Omdat algemeen $(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x}$, geldt $1 = 1^n = (p+q)^n = \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x}$.

Voorbeeld
A en B spelen een tennistoernooi van 9 wedstrijden tegen elkaar. Uit ervaring blijkt dat A ca. 6 van de 10 wedstrijden tegen B wint. Wie de meeste wedstrijden wint is winnaar van dit toernooi. Hoe groot is de kans dat B, de slechtste speler van beiden, toch wint?

Antwoord:
B wint als A hoogstens 4 successen boekt. De kans hierop is:

$$P(S_9 \leq 4) = b(0) + b(1) + b(2) + b(3) + b(4) = \binom{9}{0}(0,6)^0(0,4)^9 + \binom{9}{1}(0,6)^1(0,4)^8 + \binom{9}{2}(0,6)^2(0,4)^7 + \binom{9}{3}(0,6)^3(0,4)^6 + \binom{9}{4}(0,6)^4(0,4)^5 = (0,4)^9 + 9(0,6)(0,4)^8 + 36(0,6)^2(0,4)^7 + 84(0,6)^3(0,4)^6 + 126(0,6)^4(0,4)^5 \approx 0,2665.$$

Een kans van meer dan 25 % dus!

Het is vrij moeizaam om de kansen $\binom{n}{x} p^x q^{n-x}$ te berekenen voor grote waarden van n.

Tabelleren betekent dat men een tabel met drie ingangen (p, n en x) nodig heeft.
In de regel is men meestal geïnteresseerd in kansen van de vorm $P(S_n \leq b)$, $P(S_n \geq a)$ of $P(a \leq S_n \leq b)$.
Voor grote waarden van n zijn hiervoor uitstekende benaderingen mogelijk:
de Poisson-verdelling en de normale verdeling. Zie verderop.

Stelling
b(x) is maximaal voor $x \approx np$

Bewijs:

$P(S_n = x) = b(x) = \binom{n}{x} p^x q^{n-x}$

$b'(x)$ is moeilijk uit te rekenen. Andere ingang:

$$\frac{b(x+1)}{b(x)} = \frac{\binom{n}{x+1} p^{x+1} q^{n-x-1}}{\binom{n}{x} p^x q^{n-x}} = \frac{\frac{n!}{(x+1)!(n-x-1)!}}{\frac{n!}{x!(n-x)!}} \frac{p}{q} = \frac{n-x}{x+1} \frac{p}{q}$$

Stel $b(x)$ is max voor $x = x_m$, dan geldt

$$\frac{b(x_m)}{b(x_m - 1)} = \frac{n - x_m + 1}{x_m} \frac{p}{q} \geq 1 \quad \text{en} \quad \frac{b(x_m+1)}{b(x_m)} = \frac{n - x_m}{x_m + 1} \frac{p}{q} \leq 1$$

dus $np - x_m p + p \geq x_m q$ en $np - x_m p \leq x_m q + q$

of $np + p \geq x_m$ en $np - q \leq x_m$

zodat $\boxed{np - q \leq x_m \leq np + p}$

x_m ligt dus in een interval ter lengte van 1, dat np bevat.

Voorbeeld
9 worpen met een zuivere dobbelsteen.
Waarschijnlijkste aantal zessen: $9 \times 1/6 - 5/6 \leq x_m \leq 9 \times 1/6 + 1/6$, dus $4/6 \leq x_m \leq 10/6$, zodat $x_m = 1$.

Ander bewijs van $E(S_n) = np$:

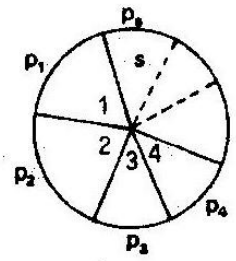
$$E(S_n) = \sum_{x=0}^{n} x\, P(S_n = x) = \sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x} = \sum_{x=1}^{n} x \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^x q^{n-x} = \sum_{\substack{y=0 \\ y=x-1}}^{n-1} \frac{n!}{y!(n-1-y)!} p^{y+1} q^{n-1-y}$$

$$= np \sum_{y=0}^{n-1} \binom{n-1}{y} p^y q^{n-1-y} = np\,(p+q)^{n-1} = np$$

9. Polynomiale verdeling (of multinomiale verdeling)

Onderstaand geluksrad wordt n keer gedraaid.



$p_1+p_2+\cdots+p_5=1$

Zij X_i het aantal keren dat de uitkomst i optreedt. Dan geldt $X_1 + X_2 + \ldots\ldots + X_s = n$.

Voor elke i heeft de toevalsvariabel X_i een binomiale verdeling. Immers, men kan de uitkomst i als succes opvatten en alle andere uitkomsten als misser.
Dan geldt:

$E(X_i) = np_i$ en $V(X_i) = np_i(1-p_i)$.

We willen nu de kans berekenen dat bij n keer draaien r_1 keer 1, r_2 keer 2, ….. , r_s keer s optreedt.

Daartoe beschouwen we eerst de uitkomst $\omega = \underbrace{11\cdots1}_{r_1}\ \underbrace{22\cdots2}_{r_2}\ \cdots\cdots\ \underbrace{55\cdots5}_{r_5}$.

Deze uitkomst heeft kans $p_1^{r_1} p_2^{r_2} \cdots p_5^{r_5}$. Alle permutaties van ω zijn echter ook gunstig! Er zijn $\frac{n!}{r_1!r_2!\cdots r_5!}$ permutaties van ω, die alle dezelfde kans hebben. Dus:

$$P(X_1=r_1 \wedge X_2=r_2 \wedge \cdots \wedge X_s=r_s) = \frac{n!}{r_1!\,r_2!\cdots r_s!}\, p_1^{r_1} p_2^{r_2} \cdots p_s^{r_s}$$

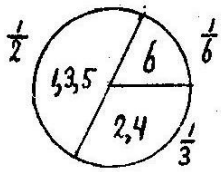
Dit is de polynomiale verdeling.

Dat er $\frac{n!}{r_1!r_2!\cdots r_5!}$ uitkomsten zijn met r_1 enen, r_2 tweeën, enz..., is ook als volgt toe te lichten.
Kies uit de n posities r_1 posities (voor de enen): dit gaat op $\binom{n}{r_1}$ manieren.
Kies uit de overblijvende $n-r_1$ posities r_2 posities (voor de tweeën): dit gaat op $\binom{n-r_1}{r_2}$ manieren.
Enz.
In totaal dus $\binom{n}{r_1}\binom{n-r_1}{r_2}\cdots = \frac{n!}{r_1!(n-r_1)!}\,\frac{(n-r_1)!}{r_2!(n-r_1-r_2)!}\cdots = \frac{n!}{r_1!\,r_2!\cdots r_5!}$ manieren.

Voorbeeld

Er wordt 6 keer met een zuivere dobbelsteen geworpen. Bereken de kans op 2 zessen en 4 oneven ogensommen.

Antwoord:



er is 6 keer gedraaid aan dit gelukswiel:

De kans op 2 zessen en 4 keer oneven is dus

$$\frac{6!}{2!4!0!}\left(\frac{1}{6}\right)^2\left(\frac{1}{2}\right)^4\left(\frac{1}{3}\right)^0 = \frac{5}{192}.$$

10. Chi-kwadraat toets

Dit is een van de nuttigste statistische toetsen.Deze hangt direct samen met de polynomiale verdeling.
We doen eerst de probleemstelling uit de doeken.
Stel dat een zeker toevalsexperiment de mogelijke uitkomsten 1, …., s heeft. We vermoeden dat deze uitkomsten resp. kans p_1, ….., p_s hebben (nulhypothese).
Wordt het experiment nu n keer herhaald, dan verwachten we dat de verschillende uitkomsten resp. np_1, ….., np_s keer zullen optreden. In werkelijkheid nemen we echter de aantallen X_1, ….., X_s waar. Deze aantallen zullen i.h.a. van de verwachte waarden verschillen. Als deze verschillen echter te groot worden, dan is er reden om de nulhypothese te verwerpen.
Hoe groot moeten de verschillen zijn om voldoende reden te hebben om de nulhypothese te verwerpen?

Voor de beantwoording van deze vraag dienen we een maat te hebben om de afwijking te meten. Een geschikte maat blijkt te zijn:

$$X^2 = \frac{(X_1-np_1)^2}{np_1} + \frac{(X_2-np_2)^2}{np_2} + \cdots\cdots + \frac{(X_s-np_s)^2}{np_s} = \sum_{i=1}^{s} \frac{(X_i-np_i)^2}{np_i}$$

Lees: chi-kwadraat
X^2 is dus een toevalsvariabele, met niet-negatieve waarden.

Wanneer is de afwijking X^2 nu te groot?
Om dat te onderzoeken hebben we de verwachtingswaarde en de spreiding (standaarddeviatie) van X^2 nodig.

Wet,

$$E(X^2) = \sum_{i=1}^{s} \frac{E((X_i-np_i)^2)}{np_i} \quad \text{wegens de Lineariteit van } E$$

$$= \sum_{i=1}^{s} \frac{V(X_i)}{np_i} \quad \text{omdat } np_i = E(X_i)$$

$$= \sum_{i=1}^{s} \frac{np_i(1-p_i)}{np_i} = \sum_{i=1}^{s} (1-p_i) = (1-p_1)+(1-p_2)+\cdots+(1-p_s) = s-1.$$

Dus bewezen:

$$E(X^2) = s-1 = f$$

Het getal f heet het aantal vrijheidsgraden, want van de aantallen X_i zijn er slechts s-1 vrij te kiezen, omdat X_1 + ….. + X_s = n.

De variantie van X^2 is echter niet zo gemakkelijk te berekenen. Uit de formule voor de X^2-toets is (met veel rekenwerk) af te leiden:

$V(X^2) \approx 2(s-1) = 2f$ voor grote waarden van n.

Hieruit volgt voor de spreiding van X^2:

$$\sigma \approx \sqrt{2f} \text{ voor grote waarden van } n$$

Als H_0 waar is, dan zal de waarde van X^2 niet erg veel verschillen van de verwachtingswaarde f. Een verschil van 2σ blijkt al erg onwaarschijnlijk te zijn.
Er is een tabel opgesteld van de mogelijke verschillen en de bijbehorende kansen.
Algemene regel: als de kans op het geconstateeerde verschil kleiner dan 5% is, dan wordt H_0 verworpen. De afwijking heet dan significant op 5%-niveau.
Hier volgt deze tabel:

X^2-TABEL

f \ P	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01
1	0,00016	0,00098	0,00393	0,01579	2,70554	3,84146	5,02389	6,63490
2	0,00201	0,00506	0,10259	0,21072	4,60517	5,99147	7,37776	9,21034
3	0,11483	0,21580	0,35185	0,58438	6,25139	7,81473	9,34840	11,3449
4	0,29711	0,48442	0,71072	1,06362	7,77944	9,48773	11,1433	13,2767
5	0,55430	0,83121	1,14548	1,61031	9,23635	11,0705	12,8325	15,0863
6	0,87209	1,23735	1,63539	2,20413	10,6446	12,5916	14,4494	16,8119
7	1,23904	1,68987	2,16735	2,83311	12,0170	14,0671	16,0128	18,4753
8	1,64648	2,17973	2,73264	3,48954	13,3616	15,5073	17,5346	20,0902
9	2,08781	2,70039	3,32511	4,16816	14,6837	16,9190	19,0228	21,6660
10	2,55821	3,24697	3,94030	4,86518	15,9871	18,3070	20,4831	23,2093
11	3,0535	3,8158	4,5748	5,5778	17,275	19,675	21,920	24,725
12	3,5706	4,4038	5,2260	6,3038	18,549	21,026	23,337	26,217
13	4,1069	5,0087	5,8919	7,0415	19,812	22,362	24,736	27,688
14	4,6604	5,6287	6,5706	7,7895	21,064	23,685	26,119	29,143
15	5,2294	6,2621	7,2604	8,5468	22,307	24,996	27,488	30,578
16	5,812	6,908	7,962	9,312	23,54	26,30	28,85	32,00
17	6,408	7,564	8,672	10,09	24,77	27,59	30,19	33,41
18	7,015	8,231	9,390	10,86	25,99	28,87	31,53	34,81
19	7,633	8,907	10,12	11,65	27,20	30,14	32,85	36,19
20	8,260	9,591	10,85	12,44	28,41	31,41	34,17	37,57
21	8,897	10,28	11,59	13,24	29,62	32,67	35,48	38,93
22	9,542	10,98	12,34	14,04	30,81	33,92	36,78	40,29
23	10,20	11,69	13,09	14,85	32,00	35,17	38,08	41,64
24	10,86	12,40	13,85	15,66	33,20	36,42	39,36	42,98
25	11,52	13,12	14,61	16,47	34,38	37,65	40,65	44,31
26	12,20	13,84	15,38	17,29	35,56	38,89	41,92	45,64
27	12,88	14,57	16,15	18,11	36,74	40,11	43,19	46,96
28	13,56	15,31	16,93	18,94	37,92	41,34	44,46	48,28
29	14,26	16,05	17,71	19,77	39,09	42,56	45,72	49,59
30	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89

Toelichting op deze tabel.
De 5e rij betekent: bij 5 vrijheidsgraden (f = 5) geldt
$P(X^2 \geq 0{,}55430) = 0{,}99$ $P(X^2 \geq 0{,}83121) = 0{,}975$ ... enz.

Voorbeeld
Van een zekere plant is bekend, uit de wetten van Mendel, dat 4 typen voorkomen, in de verhouding 9 : 3 : 3 : 1.
In een steekproef van 240 van deze planten komen de 4 typen voor in resp. de aantallen 120, 40, 55 en 25.
Is deze afwijking van de theoretische aantallen significant op 1 % - nivo?

Antwoord:
$X^2 = 15^2/135 + 5^2/45 + 10^2/45 + 10^2/15 \approx 11{,}1$ en f = 3, dus volgens de tabel geldt net niet $P(X^2 \geq 11{,}3449) = 0{,}01$.
Dus antwoord: net niet juist.

Voorbeeld
Een dobbelsteen wordt getoetst op zuiverheid. De dobbelsteen wordt 120 keer geworpen, met als resultaat:

Mogelijke ogensom i	1	2	3	4	5	6
Waargenomen aantal keren Xi	15	21	25	19	14	26
Verwachte aantal keren npi	20	20	20	20	20	20

Bestaat er nu reden om aan de zuiverheid van deze dobbelsteen te twijfelen?

$\chi^2 = \frac{5^2}{20} + \frac{1^2}{20} + \frac{5^2}{20} + \frac{1^2}{20} + \frac{6^2}{20} + \frac{6^2}{20} = \frac{124}{20} = 6,2$ en $f = 6-1 = 5$. In de tabel lezen we: $P(\chi^2 \geq 1,61031) = 0,90$.
Er is dus geen reden om H_0 (de dobbelsteen is zuiver) te verwerpen, tenminste niet op grond van deze resultaten.

Voorbeeld

A leent zijn zuivere dobbelsteen aan B. De volgende dag beweert B in 60 worpen met deze dobbelsteen 10 keer 1, 10 keer 2, 10 keer 3, 9 keer 4, 9 keer 5 en 12 keer 6 te hebben gegooid.
A twijfelt, uit nieuwsgierigheid, snel aan de χ^2-toets uit:
$\chi^2 = \frac{0^2}{10} + \frac{0^2}{10} + \frac{0^2}{10} + \frac{1^2}{10} + \frac{1^2}{10} + \frac{2^2}{10} = 0,6$ en $f = 6-1 = 5$. Dus $\chi^2 < f$. In de tabel lezen we: $P(\chi^2 < 0,83121) = 1 - P(\chi^2 \geq 0,83121) = 1 - 0,975 = 0,025$. A heeft dus reden om H_0 (de dobbelsteen is zuiver) te verwerpen. Maar hij weet dat de dobbelsteen zuiver is! Wat moet A nu concluderen?
Wel, bij de probleemstelling aan het begin van deze par., hebben we meestal de hypothese H_0 nog een veronderstelling gemaakt. Nl. dat het experiment n keer herhaald wordt. Dus hier moet iets mis mee zijn. B heeft kennelijk geen 60 keer geworpen, maar de resultaten gewoon uit zijn duim gezogen. Met als gevolg dat de resultaten te mooi werden.

Een te kleine waarde van χ^2 kan dus net zo verdacht zijn als een te grote.
Bij een te kleine waarde van χ^2 moet je er rekening mee houden dat iemand de gegevens gefabriceerd heeft, om daarmee iets te bewijzen.

Toepassing van de X^2-toets op de viervakkentabel i.p.v. de exacte toets van Fisher,
Voorbeeld

Bij 300 personen werd een onderzoek verricht naar mogelijk verband tussen geslacht en haarkleur. Na telling bleek:

	licht	donker
mannen	72	128
vrouwen	48	52

Onze hypotheses zijn

H_0: tussen haarkleur en geslacht bestaat geen verband

H_1: bij mannen komt een donkere haarkleur meer voor dan bij vrouwen.

De exacte test van Fisher zou als volgt verlopen.

Stel dat H_0 waar is. Dan is de kans, om in een steekproef van 200 mannen minstens 128 donkerharige aan te treffen, gelijk aan:

$$\alpha = \frac{\binom{120}{72}\binom{180}{128} + \binom{120}{71}\binom{180}{129} + \ldots\ldots\ldots\ldots + \binom{120}{20}\binom{180}{180}}{\binom{300}{200}}$$

Ufff, probeer α maar eens uit te rekenen!

Gelukkig blijkt de χ^2-toets uitkomst te bieden. Dit gaat als volgt.

Onder de aanname H_0 zouden we verwacht hebben dat de verhouding 120 (licht) : 180 (donker) zowel voor de mannen als voor de vrouwen zou gelden. We hadden dan deze viervakkentafel verwacht:

	licht	donker
mannen	80	120
vrouwen	40	60

Het verschil tussen aangetroffen en verwachte aantallen meten we met χ^2.

$$\chi^2 = \frac{8^2}{80} + \frac{8^2}{120} + \frac{8^2}{40} + \frac{8^2}{60} = 4.$$

De X^2-toets mag nu worden toegepast (geen bewijs), mits we voor het aantal vrijheidsgraden f niet 3 maar 1 nemen (geen bewijs). De tabel geeft: $P(X^2 \geq 3{,}84146) = 0{,}05$. De afwijking is dus significant op 5%-nivo. We hebben dus reden om H_0 te verwerpen en H_1 te accepteren. Helaas, haarkleur is niet afhankelijk van het geslacht! We moeten dus naar een andere verklaring zoeken. Bijv. dat een deel van onderzochte personen een haarspoeling heeft gebruikt.

N.B.
Voor toepassing van de X^2-toets op de viervakkentabel is vereist dat $n \geq 30$ en dat de verwachte aantallen minstens 5 zijn (geen bewijs).
De exacte toets van Fisher is terughoudender dan de Chi-kwadraat toets in het benoemen of er een significant verschil is tussen beide groepen.

<u>Salk-vaccin</u>

Jonas Salk ontdekte in 1953 het eerste polio-vaccin. In 1954 werd in de VS een grootscheeps onderzoek van dit vaccin tegen kinderverlamming uitgevoerd. Er werden

401974 kinderen ingeënt, de ene helft (proefgroep) met het Salk-vaccin, de andere helft (placebogroep of controlegroep) met een zoutoplossing.
De resultaten waren opzienbarend zoals bijgaande tabel laat zien.

	polio	geen polio
proefgroep	33	200712
placebo-groep	115	201114

Zonder berekeningen kan hier direct worden vastgesteld dat dit vaccin uitermate effectief is. Als we toch nog even de chi-2 toets loslaten op deze viervakkentabel, dan vinden we $\chi^2 = 45{,}45$. Met $f = 1$ geeft de chi-2 tabel $P(\chi^2 \geq 6{,}63) = 0{,}01$ dus de gevraagde kans is veel kleiner dan 0,01.

11. Benadering van de binomiale verdeling door de Poisson-verdeling in bep. gevallen

Binomiale verdeling:
Experiment: n herhalingen van hetzelfde experiment waarbij per keer de succeskans p is.
Dus elke uitkomst is een rij van enen (successen) en nullen (missers).
Zij S_n het aantal successen na deze n herhalingen.

Dan geldt: $E(S_n) = np$, $V(S_n) = npq$ en $\sigma(S_n) = \sqrt{(npq)}$. Plus de formule:

$$P(S_n = x) = b(x) = \binom{n}{x} p^x q^{n-x} \quad \text{met } x = 0, 1, \ldots, n.$$

Deze formule is in bepaalde gevallen echter niet erg handig.
Voorbeeld
De grote jubileumdag.
Op 24 januari 2023 hoopt een groot bedrijf zijn 100-jarig bestaan te vieren. De directie besluit voor alle kinderen van werknemers, die op de jubileumdag geboren worden, een spaarrekening van 1000 euro te openen.
Bekend is dat binnen het bedrijf gemiddeld 730 kinderen per jaar worden geboren, dus gemiddeld 2 per dag. Men moet dus rekening houden met 2 gelukkigen. Om uitschieters een slag voor te zijn, wordt 5.000 euro gereserveerd.
Hoe groot is de kans dat dit bedrag toch niet toereikend is?

730 kinderen per jaar betekent 730 kinderen in 365 dagen. Hoe groot is de kans dat meer dan 5 daarvan op 24 januari 2023 worden geboren? Anders gezegd: verdeel 730 knikkers aselect over 365 verschillende vazen en bereken de kans dat in de 24^e vaas meer dan 5 knikkers komen. Dus n = 730 en p = 1/365. Dan $P(S_n>5) = 1 - P(S_n\leq 5) = 1 - b(0) - b(1) - b(2) - b(3) - b(4) - b(5)$. Dit wordt een moeizame berekening met bovenstaande formule!

Het is echter mogelijk een benaderingsformule voor b(x) af te leiden die wel gemakkelijk te hanteren is. Deze benadering geldt alleen voor kleine waarden van p.
Hier volgt de afleiding.

Stel p is erg klein, nl. $0<p<0,1$.

De verwachtingswaarde E = np kan dan ook nog vrij klein zijn, zelfs voor grotere waarden van n.

We weten:

$$b(0) = \binom{n}{0} p^0 (1-p)^{n-0} = (1-p)^n.$$

Omdat $\lim_{h\to 0} (1+h)^{\frac{1}{h}} = e$ (in de analyse bewezen), geldt $(1+h)^{\frac{1}{h}} \approx e$ voor kleine waarden van h.

Dus $(1-p)^n = \left\{(1-p)^{\frac{1}{-p}}\right\}^{np} = \left\{(1-p)^{\frac{1}{-p}}\right\}^{-E} \approx e^{-E}$

mits E vrij klein is.

We hebben nu gevonden:

$$b(0) \approx e^{-E}$$

Nu geldt $\frac{b(x+1)}{b(x)} = \frac{\binom{n}{x+1} p^{x+1}(1-p)^{n-x-1}}{\binom{n}{x} p^x (1-p)^{n-x}} = \frac{\frac{n(n-1)\ldots(n-x)}{(x+1)!}}{\frac{n(n-1)\ldots(n-x+1)}{x!}} \frac{p}{1-p} = \frac{n-x}{x+1}\frac{p}{1-p} = \frac{1-\frac{x}{n}}{x+1}\frac{np}{1-p} = \frac{1-\frac{x}{n}}{x+1}\frac{E}{1-p}$

$$\approx \frac{1}{x+1}\frac{E}{1} \quad \text{als } \frac{x}{n} \text{ erg klein is.}$$

Hieruit volgt:

$$\frac{b(x+1)}{b(x)} \approx \frac{E}{x+1}, \text{ dus } b(x+1) \approx \frac{E}{x+1} b(x), \text{ mits } 0<p<0,1 \text{ en } E, \frac{x}{n} \text{ klein.}$$

Zo vinden we: $b(1) \approx \frac{E}{1} b(0) \approx \frac{E}{1} e^{-E}$

$$b(2) \approx \frac{E}{2} b(1) \approx \frac{E}{2}\frac{E}{1} e^{-E} = \frac{E^2}{2!} e^{-E}$$

$$b(3) \approx \frac{E}{3} b(2) \approx \frac{E}{3}\frac{E^2}{2!} e^{-E} = \frac{E^3}{3!} e^{-E}$$

Algemeen geldt:

$$b(x) \approx \frac{E^x}{x!} e^{-E} \text{ als } 0<p<0,1 \text{ en } E, \frac{x}{n} \text{ klein.}$$

Deze benadering voor de binomiale verdeling wordt de Poisson-verdeling genoemd (naar Poisson, 1781–1840).

Voor de variantie geldt: $V = np(1-p) \approx np = E$.

Dus:

$$V \approx E = np.$$

Voor deze benadering geldt $\sum b(x) = \sum \frac{E^x}{x!} e^{-E} = e^{-E}e^{E} = 1$, dus deze benadering is zelf ook een kansverdeling.

Deze benaderingsformule is ook zo af te leiden:

$$P(S_N = x) = \binom{N}{x} p^x (1-p)^{N-x} \qquad x = 0, 1, \ldots, N$$

$$= \frac{N!}{x!(N-x)!} p^x (1-p)^{N-x}$$

$m = Np$: $p = \frac{m}{N}$, dus $P(S_N = x) = \frac{N!}{x!(N-x)!} \left(\frac{m}{N}\right)^x \left(1 - \frac{m}{N}\right)^{N-x}$

$$= \frac{N(N-1)\cdots(N-x+1)}{x!} \frac{m^x}{N^x} \left(1 - \frac{m}{N}\right)^N \left(1 - \frac{m}{N}\right)^{-x} = \frac{N(N-1)\cdots(N-x+1)}{N^x} \frac{m^x}{x!} \left(1 - \frac{m}{N}\right)^N \left(1 - \frac{m}{N}\right)^{-x}$$

$N \to \infty$: $P(S_N = x) = 1 \cdot \frac{m^x}{x!} \left(1 - \frac{m}{N}\right)^N \left(1 - \frac{m}{N}\right)^{-x} \to \frac{m^x}{x!} \left(1 - \frac{m}{N}\right)^N \cdot 1 \to \frac{m^x}{x!} e^{-m}$

de bekende formule!

Nu de berekening bij de grote jubileumdag.
De benadering van Poisson, met p = 1/365, E = 730x1/365 = 2 en x/n ≤ 5/730, geeft:

$$1 - b(0) - b(1) - b(2) - b(3) - b(4) - b(5) \approx 1 - \frac{2^0}{0!} e^{-2} - \frac{2^1}{1!} e^{-2} - \frac{2^2}{2!} e^{-2} - \frac{2^3}{3!} e^{-2} - \frac{2^4}{4!} e^{-2} - \frac{2^5}{5!} e^{-2} = 1 - \frac{109}{15} e^{-2} \approx 0{,}0168.$$

Conclusie: de kans op een onaangename verrassing is dus zeer klein (minder dan 2%). Men hoeft daar geen rekening mee te houden.

Echte berekening volgens de binomiale verdeling:
$P(S_n>5) = 1 – P(S_n≤5) = 1 – b(0) – b(1) – b(2) – b(3) – b(4) – b(5) = 0{,}015$ m.b.v. een computerprogramma (op internet te vinden).
De Poisson-benadering is dus heel nauwkeurig!

Voorbeeld
In een bepaalde stad vinden gemiddeld 2 zelfmoorden per week plaats. Dan is er een week met 5 zelfmoorden. Is dit nog als “normaal” te beschouwen?

Antwoord:
Gemiddeld 2 zelfmoorden per week is waarschijnlijk gebaseerd op jaarlijkse tellingen die geleid hebben tot de conclusie dat er gemiddeld 104 zelfmoorden per jaar (52 weken) gebeuren. Meer weten we misschien niet. We nemen gemakshalve maar aan dat de 104 zelfmoorden aselect verdeeld zijn over de 52 weken. Dan zou het vergelijkbaar kunnen zijn met het aselect verdelen van 104 knikkers over 52 verschillende vazen. D.w.z. dat het om 104 experimenten gaat die onafhankelijk van elkaar zijn. Nu blijken 5 knikkers in vaas i terecht te komen. Hoe normaal is dit?
Dus het komt dan neer op een binomiaal experiment met n = 104 en p = 1/52, waarbij p de kans is dat een knikker in vaas i komt. Gevraagd wordt dan de kans dat 5 of meer knikkers in vaas i terecht komen ofwel $P(S_n \geq 5)$.
Echte berekening volgens de binomiale verdeling (m.b.v een computerprogramma) levert:
$P(S_n \geq 5) = 1 – P(S_n) = 1 – b(0) – b(1) – b(2) – b(3) – b(4) = 0{,}053$.
Dit is net niet significant!
De voorwaarden laten echter toe de benadering met de Poisson-verdeling toe te passen.
Deze levert, met E = np = 2:

$P(S_n \geq 5) = 1 - P(S_n < 5) = 1- b(0) - b(1) - b(2) - b(3) - b(4) \approx 1 - 7^{e-2} = 0{,}053$.
Dus we zien opnieuw dat de benadering van Poisson heel nauwkeurig is.

Een foute redenering is de volgende.
2 zelfmoorden per week betekent 2 zelfmoorden per 7 dagen. Maar dit komt niet neer op 2 knikkers aselect verdelen over 7 verschillende vazen, want dan kunnen er nooit meer dan 2 knikkers in dezelfde vaas komen.

12. Benadering van de binomiale verdeling door de normale verdeling in bep. gevallen

Binomiale verdeling:
Experiment: n herhalingen van hetzelfde experiment waarbij per keer de succeskans p is. Dus elke uitkomst is een rij van enen (successen) en nullen (missers).
Zij S_n het aantal successen na deze n herhalingen.

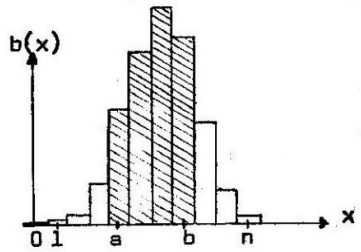
Dan geldt: $E(S_n) = np$, $V(S_n) = npq$ en $\sigma(S_n) = \sqrt{(npq)}$.
En:

$$P(S_n = x) = b(x) = \binom{n}{x} p^x q^{n-x} \quad \text{met } x = 0, 1, \ldots, n.$$

Als we echter $P(a \le S_n \le b)$ moeten berekenen, dan wordt dat heel lastig. Zeker als n groot is en a en b ook nog erg verschillen. Ook hiervoor is een goede benadering mogelijk die wel goed hanteerbaar is. Deze benadering wordt mogelijk door de som $P(a \le S_n \le b) = \sum b(x)$ te vervangen door een integraal. Een integraal die niet exact te berekenen is maar wel goed te benaderen is.

Hier volgt de afleiding.

De waarde van $P(a \le S_n \le b) = \sum_{x=a}^{b} b(x)$ is juist de oppervlakte van het hieronder gearceerde gebied:



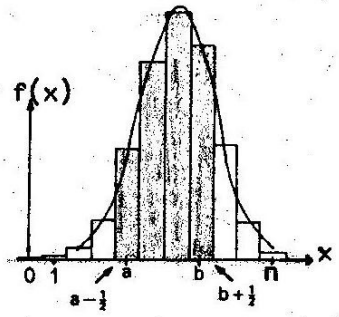
De functie $b(x)$ is alleen voor niet-negatieve gehele waarden van x gedefinieerd.
We proberen nu een functie $f(x)$ te vinden die voor alle niet-negatieve reële waarden van x gedefinieerd is en waarvoor geldt:

$b(x) \approx f(x)$, als x geheel is.

Een adequate kandidaat voor de functie f is te vinden m.b.v. de formule van Stirling:

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n} \quad \text{voor alle } n$$

We krijgen dan dit plaatje:



Meetkundig is nu direct duidelijk dat $P(a \leq S_n \leq b) = \sum_{x=a}^{b} b(x) \approx \int_{a-\frac{1}{2}}^{b+\frac{1}{2}} f(x)\,dx$.

Welnu, voor de functie f blijken we te kunnen nemen (geen bewijs):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-np}{\sigma}\right)^2}, \text{ met } \sigma^2 = V(S_n) = npq.$$

Deze functie $f(x)$ is een goede benadering voor $b(x)$, mits $npq > 9$ (geen bewijs).
De functie f wordt de normale kansdichtheid genoemd. Het woord dichtheid geeft aan dat x niet geheel hoeft te zijn: de discrete verdeling

x	.....
b(x)	.....

is vervangen door de continue verdeling

x	----
f(x)	----

Om dus $P(a \leq S_n \leq b)$ te benaderen, moeten we de bepaalde integraal $\int_{a-\frac{1}{2}}^{b+\frac{1}{2}} f(x)\,dx$ uitrekenen.
Daartoe substitueren we eerst $\frac{x-np}{\sigma} = u$. De integraal gaat dan over in:

$$\int_{\frac{a-\frac{1}{2}-np}{\sigma}}^{\frac{b+\frac{1}{2}-np}{\sigma}} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}u^2}\,\sigma\,du = \int_{\frac{a-\frac{1}{2}-np}{\sigma}}^{\frac{b+\frac{1}{2}-np}{\sigma}} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}u^2}\,du = \Phi\left(\frac{b+\frac{1}{2}-np}{\sigma}\right) - \Phi\left(\frac{a-\frac{1}{2}-np}{\sigma}\right)$$

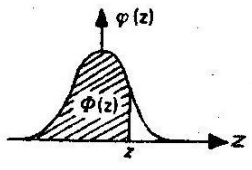
waarbij $\Phi'(z) = \varphi(z) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}z^2}$.

Helaas, zo'n primitieve functie Φ is er niet.

We kunnen $\Phi(z)$ wel opvatten als een oppervlakte en voor het berekenen van oppervlakten bestaan goede benaderingstechnieken. Op deze wijze zijn de functiewaarden van $\Phi(z)$ te tabelleren.

TABEL

van de functie $\Phi(z)$.



Voor $z<0$ kan men gebruik maken van:

$$\Phi(-z) = 1 - \Phi(z).$$

z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$
.00	.50000	.80	.78814	1.60	.94520	2.40	.99180	3.20	.99931
.02	.50798	.82	.79389	1.62	.94738	2.42	.99224	3.22	.99936
.04	.51595	.84	.79955	1.64	.94950	2.44	.99266	3.24	.99940
.06	.52392	.86	.80511	1.66	.95154	2.46	.99305	3.26	.99944
.08	.53188	.88	.81057	1.68	.95352	2.48	.99343	3.28	.99948
.10	.53983	.90	.81594	1.70	.95543	2.50	.99379	3.30	.99952
.12	.54776	.92	.82121	1.72	.95728	2.52	.99413	3.32	.99955
.14	.55567	.94	.82639	1.74	.95907	2.54	.99446	3.34	.99958
.16	.56356	.96	.83147	1.76	.96080	2.56	.99477	3.36	.99961
.18	.57142	.98	.83646	1.78	.96246	2.58	.99506	3.38	.99964
.20	.57926	1.00	.84134	1.80	.96407	2.60	.99534	3.40	.99966
.22	.58706	1.02	.84614	1.82	.96562	2.62	.99560	3.42	.99969
.24	.59483	1.04	.85083	1.84	.96712	2.64	.99585	3.44	.99971
.26	.60257	1.06	.85543	1.86	.96856	2.66	.99609	3.46	.99973
.28	.61026	1.08	.85993	1.88	.96995	2.68	.99632	3.48	.99975
.30	.61791	1.10	.86433	1.90	.97128	2.70	.99653	3.50	.99977
.32	.62552	1.12	.86864	1.92	.97257	2.72	.99674	3.52	.99978
.34	.63307	1.14	.87286	1.94	.97381	2.74	.99693	3.54	.99980
.36	.64058	1.16	.87698	1.96	.97500	2.76	.99711	3.56	.99981
.38	.64803	1.18	.88100	1.98	.97615	2.78	.99728	3.58	.99983
.40	.65542	1.20	.88493	2.00	.97725	2.80	.99744	3.60	.99984
.42	.66276	1.22	.88877	2.02	.97831	2.82	.99760	3.62	.99985
.44	.67003	1.24	.89251	2.04	.97932	2.84	.99774	3.64	.99986
.46	.67724	1.26	.89617	2.06	.98030	2.86	.99788	3.66	.99987
.48	.68439	1.28	.89973	2.08	.98124	2.88	.99801	3.68	.99988
.50	.69146	1.30	.90320	2.10	.98214	2.90	.99813	3.70	.99989
.52	.69847	1.32	.90658	2.12	.98300	2.92	.99825	3.72	.99990
.54	.70540	1.34	.90988	2.14	.98382	2.94	.99836	3.74	.99991
.56	.71226	1.36	.91308	2.16	.98461	2.96	.99846	3.76	.99992
.58	.71904	1.38	.91621	2.18	.98537	2.98	.99856	3.78	.99992
.60	.72575	1.40	.91924	2.20	.98610	3.00	.99865	3.80	.99993
.62	.73237	1.42	.92220	2.22	.98679	3.02	.99874	3.82	.99993
.64	.73891	1.44	.92507	2.24	.98745	3.04	.99882	3.84	.99994
.66	.74537	1.46	.92785	2.26	.98809	3.06	.99889	3.86	.99994
.68	.75175	1.48	.93056	2.28	.98870	3.08	.99896	3.88	.99995
.70	.75804	1.50	.93319	2.30	.98928	3.10	.99903	3.90	.99995
.72	.76424	1.52	.93574	2.32	.98983	3.12	.99910	3.92	.99996
.74	.77035	1.54	.93822	2.34	.99036	3.14	.99916	3.94	.99996
.76	.77637	1.56	.94062	2.36	.99086	3.16	.99921	3.96	.99996
.78	.78230	1.58	.94295	2.38	.99134	3.18	.99926	3.98	.99997

De totale oppervlakte van de “klok” is $\int_{\infty}^{\infty} \varphi(z)dz$ en de waarde van deze bekende integraal is 1. Vanwege de symmetrie van de klok geldt $\Phi(-z) = 1 - \Phi(z) =$

Dus we hebben afgeleid:

$$P(a \leq S_n \leq b) = \sum_{x=a}^{b} \binom{n}{x} p^x q^{n-x} \approx \Phi\left(\frac{b+\frac{1}{2}-np}{\sigma}\right) - \Phi\left(\frac{a-\frac{1}{2}-np}{\sigma}\right)$$

mits $npq > 9$.

De functie Φ wordt normale verdeling of Gauss-verdeling genoemd.

Voorbeeld
Er wordt 6000 keer met een zuivere dobbelsteen geworpen. Hoe groot is de kans dat het aantal zessen tussen 900 en 1100 ligt?

Antwoord:
n = 6000, p = 1/6 en σ = √(6000.1/6.5/6) ≈28,868. dus $P(900 \le S_n \le 1100) \approx \Phi(100{,}5/\sigma) - \Phi(-100{,}5/\sigma) = 2\Phi(100{,}5/\sigma) - 1 \approx 2\Phi(3{,}48) - 1 \approx 2x0{,}99975 - 1 \approx 0{,}9995$.
Dus de kans dat het aantal zessen meer dan 100 afwijkt van het verwachte aantal 1000 is verwaarloosbaar klein!

Hoe groot is de kans dat het aantal zessen tussen 950 en 1050 ligt?
$P(950 \le S_n \le 1050) \approx \Phi(50{,}5/\sigma) - \Phi(-50{,}5/\sigma) = 2\Phi(50{,}5/\sigma) - 1 = 2\Phi(1{,}749) - 1 \approx 2x0{,}96 - 1 \approx 0{,}92$.
En zonder continuïteitscorrectie dan?
Dan wordt de kans $2\Phi(50/\sigma) - 1 \approx 2\Phi(1{,}732) - 1 = 2 \times 0{,}964 - 1 = 0{,}928$.

Binomiale verdeling: afwijking tussen S_n en np

Gevolgen voor de benaderingsformule $P(a \le S_n \le b) \approx \Phi(b+½-np\,/\,\sigma) - \Phi(a-½-np\,/\,\sigma)$ als npq > 9.
Zie onderstaande afleiding.

$|S_n - np| < n\sigma$ oftwel $np - n\sigma < S_n < np + n\sigma$
Kans hierop: $\Phi\left(\frac{np+n\sigma+\frac{1}{2}-np}{\sigma}\right) - \Phi\left(\frac{np-n\sigma-\frac{1}{2}-np}{\sigma}\right) = 2\Phi\left(\frac{n\sigma+\frac{1}{2}}{\sigma}\right) - 1 = 2\Phi\left(n+\frac{1}{2\sigma}\right) - 1 \approx \boxed{2\Phi(n)-1}$ voor grote σ

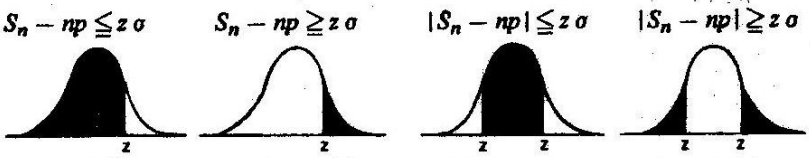
$|S_n - np| \ge n\sigma$
Complementaire gebeurtenis: $|S_n - np| < n\sigma$ oftwel $|S_n - np| < n\sigma - 1$ oftwel $np - n\sigma + 1 < S_n < np + n\sigma - 1$
Kans hierop: $\Phi\left(\frac{np+n\sigma-1+\frac{1}{2}-np}{\sigma}\right) - \Phi\left(\frac{np-n\sigma+1-\frac{1}{2}-np}{\sigma}\right) = \Phi\left(\frac{n\sigma-\frac{1}{2}}{\sigma}\right) - \Phi\left(\frac{-n\sigma+\frac{1}{2}}{\sigma}\right) = 2\Phi\left(\frac{n\sigma-\frac{1}{2}}{\sigma}\right) - 1 = 2\Phi\left(n-\frac{1}{2\sigma}\right) - 1 \approx 2\Phi(n)-1$
Dus de kans op $|S_n - np| \ge n\sigma$: $1 - \{2\Phi(n)-1\} = \boxed{2-2\Phi(n)}$ voor grote σ
voor grote σ

$S_n - np < n\sigma$
oftwel $0 < S_n < np + n\sigma$
Kans hierop: $\Phi\left(\frac{np+n\sigma+\frac{1}{2}-np}{\sigma}\right) - \Phi\left(\frac{0-\frac{1}{2}-np}{\sigma}\right) = \Phi\left(n+\frac{1}{2\sigma}\right) - \Phi\left(\frac{-1}{2\sigma} - \frac{\sigma}{q}\right)$; voor grote σ ligt $\frac{-\sigma}{q}$ vrij ver links van 0, zodat $\Phi\left(\frac{-1}{2\sigma} - \frac{\sigma}{q}\right)$ verwaarloosbaar klein is (zie de fig. bij de tabel van Φ).
Dus de kans op $S_n - np < n\sigma$ is voor grote σ bij benadering $\boxed{\Phi(n)}$

$S_n - np \ge n\sigma$
Complementaire gebeurtenis: $S_n - np < n\sigma$ oftwel $S_n - np < n\sigma - 1$ oftwel $0 < S_n < np + n\sigma - 1$
Kans hierop: $\Phi\left(\frac{np+n\sigma-1+\frac{1}{2}-np}{\sigma}\right) - \Phi\left(\frac{0-\frac{1}{2}-np}{\sigma}\right) = \Phi\left(n-\frac{1}{2\sigma}\right) - \Phi\left(\frac{-1}{2\sigma} - \frac{\sigma}{q}\right) \approx \Phi(n) - 0 = \Phi(n)$.
De kans op $S_n - np \ge n\sigma$ is dus omgeveer $\boxed{1-\Phi(n)}$ voor grote σ

Hier volgt een overzicht van de resultaten, tesamen met een handige tabel:

TABEL van de benaderde kansen op de volgende vier gebeurtenissen, voor G niet te klein:



z	$\Phi(z)$	$1-\Phi(z)$	$2\Phi(z)-1$	$2-2\Phi(z)$
0.0	.500	.500	.0000	1.0000
0.1	.540	.460	.0797	.9203
0.2	.579	.421	.159	.841
0.3	.618	.382	.236	.764
0.4	.655	.345	.311	.689
0.5	.691	.309	.383	.617
0.6	.726	.274	.451	.549
0.7	.758	.242	.516	.484
0.8	.788	.212	.576	.424
0.9	.816	.184	.632	.368
1.0	.841	.159	.683	.317
1.1	.864	.136	.729	.271
1.2	.885	.115	.770	.230
1.3	.9032	.0968	.806	.194
1.4	.9192	.0808	.838	.162
1.5	.9332	.0668	.866	.134
1.6	.9452	.0548	.890	.110
1.7	.9554	.0446	.9109	.0891
1.8	.9641	.0359	.9281	.0719
1.9	.9713	.0287	.9425	.0575
2.0	.9772	.0228	.9545	.0455
2.1	.9821	.0179	.9643	.0357
2.2	.9861	.0139	.9722	.0278
2.3	.9893	.0107	.9786	.0214
2.4	.99180	.00820	.9836	.0164
2.5	.99379	.00621	.9876	.0124
2.6	.99534	.00466	.99068	.00932
2.7	.99653	.00347	.99307	.00693
2.8	.99744	.00256	.99489	.00511
2.9	.99813	.00187	.99627	.00373
3.0	.99865	.00135	.99730	.00270
3.1	.999032	.000968	.99806	.00194
3.2	.999313	.000687	.99863	.00137
3.3	.999517	.000483	.999033	.000967
3.4	.999663	.000337	.999326	.000674
3.5	.999767	.000233	.999535	.000465
3.6	.999841	.000159	.999682	.000318
3.7	.999892	.000108	.999784	.000216
3.8	.9999277	.0000723	.999855	.000145
3.9	.9999519	.0000481	.9999038	.0000962
4.0	.9999683	.0000317	.9999367	.0000633

Voorbeeld

Hoe groot is de kans bij 600 worpen met een zuivere dobbelsteen minstens 110 keer een zes te gooien?

Antwoord:

n = 600, p = 1/6 en $\sigma = \sqrt{(600.1/6.5/6)} = 10\sqrt{\frac{5}{6}}$

$P(S_n \geq 110) = P(S_n - np \geq 10) = P(S_n - np \geq 1{,}1\ \sigma) = 1 - \Phi(1{,}1) = 0{,}136$, dus > 5%.

Deze kans is dus niet significant klein!

Nog een voorbeeld

Bij een telefonisch meldpunt komen wekelijks gemiddeld 50 schademeldingen binnen. Dan is er een week met 60 schademeldingen. Is dit nog als normaal te beschouwen?
Niet een van de meest dringende problemen, maar wel een voorbeeld van dat hier op drie manieren een kansberekening bij gegeven kan worden.

Allereerst, 50 schademeldingen per week kan de resultante zijn van dat jaarlijks gemiddeld 2600 schademeldingen binnenkomen. Aannemend dat deze nagenoeg onafhankelijk van elkaar zijn, is dit vergelijkbaar met de aselecte verdeling van 2600 knikkers over 52 verschillende vazen. Dus er is sprake van een binomiale verdeling met n = 2600, p = 1/52 ≈ 0,01923 en μ = E = np = 50. De vraag is dan wat de uitkomst van $P(S_n \geq 60)$ is.
De binomiale verdeling geeft, m.b.v. een computerprogramma, als uitkomst 0,09.
De benadering met de Poisson-verdeling mag ook worden toegepast omdat aan de voorwaarden hiervoor is voldaan. Dan geldt $P(S_n \geq 60) = 1 - P(S_n < 60) = 1 - b(0) - b(1) - \ldots\ldots\ldots - b(59) \approx 1 - e^{-50}\,(50^0/0! - 50^1/1! - 50^2/2! - \ldots\ldots\ldots\ldots - 50^{59}/59!)$. Ook hiervoor is een computerprogramma nodig en dit geeft als uitkomst 0,09.
Handiger is hier echter de benadering met de normale verdeling. Dit mag, want er is voldaan aan de voorwaarde npq = np(1–p) > 9. De uitkomst van $P(S_n \geq 60) = P(S_n - np \geq 10) = P(S_n - np \geq 1{,}43\sigma) \approx 1 - \Phi(1{,}43) \approx 0{,}08$.

Nog een interessante beschouwing: De macht van een vastberaden minderheid
Een kleine vastberaden minderheid kan in een onverschillige populatie een geweldige invloed uitoefenen.
Dit laten we zien aan de hand van een tweetal voorbeelden.

1)
Een vereniging bestaat uit 25 leden. 5 leden zijn voorstander van een bepaald voorstel, de overige 20 leden maakt het niets uit. Zodra 8 van hen voorstemmen, is het voorstel erdoor.
De kans hierop is volgens de binomiale verdeling gelijk aan:

$$\sum_{x=8}^{20}\binom{20}{x}\left(\tfrac{1}{2}\right)^{x}\left(\tfrac{1}{2}\right)^{20-x} = \sum_{x=8}^{20}\binom{20}{x}\left(\tfrac{1}{2}\right)^{20} = 0{,}89.$$

De aanname hierbij is dat de 20 onverschillige leden aselect stemmen. Dat is natuurlijk niet zeker.
De benadering van deze kans m.b.v. de normale verdeling is niet toegestaan, omdat niet voldaan is aan de voorwaarde npq > 9, want npq = 20 x ½ x ½ = 5.

2)
Gegeven: een populatie van 15 miljoen moet stemmen over een bepaald voorstel.
Stel dat minderheid van n personen resoluut voorstander is en dat de rest van de populatie geen voorkeur heeft en aselect stemt.
Hoe groot moet n zijn om het voorstel aangenomen te krijgen?

Oplossing:
Er zijn n resoluten en $15 \times 10^6 - n$ onverschilligen die aselect stemmen.
Zij S_n het aantal voorstemmers onder de onverschilligen. Dan $\mu = E(S_n) = \tfrac{1}{2} \times 15 \times 10^6 - \tfrac{1}{2} n$ en $\sigma = \sqrt{\mu x \tfrac{1}{2}} = \sqrt{3{,}75 \times 10^6 - \tfrac{1}{4}n}$

Hoe groot is nu de kans dat $n + S_n > 7{,}5 \times 10^6$ ofwel dat $S_n - (7{,}5 \times 10^6 - \tfrac{1}{2} n) > -\tfrac{1}{2} n$

$= \dfrac{-\tfrac{1}{2}\,n}{\sqrt{3{,}75\times10^6-\tfrac{1}{4}n}}\ \sigma = -z\,\sigma$?

De gevraagde kans is 1- Φ(-z) = Φ(z).
Voor z = 1,65 is deze kans al rond 95%.

Dus n nog op lossen uit $\frac{½\ n}{\sqrt{3{,}75\times10^6-\frac{1}{4}n}}$ = 1,65.

Dit leidt tot de vierkantsvergelijking $(n+1{,}36)^2 = 1{,}36^2 + 40{,}8 \times 10^6$. Hieruit volgt n ≈ 6385 á 6390.

12.1. Betrouwbaarheidsinterval van een schatting

We weten dat de kans op de gebeurtenis $|S_n - np| \leq z\sigma$ nagenoeg gelijk is aan $2\Phi(z) - 1$.
Omwerken geeft $|S_n / n - p| \leq z\,\sigma/n = z\sqrt{(npq)} / n = z\sqrt{(pq/n)}$.
S_n /n is een schatting (benadering) van p; we schrijven hiervoor $\hat{p}$.
Dus $|\hat{p} - p| \leq z\sqrt{(pq/n)}$.
Als we nu in het rechterlid p en q vervangen door hun schattingen $\hat{p}$ en $\hat{q}$, dan geeft dit slechts een klein verschil, omdat door n wordt gedeeld.
De ongelijkheid gaat dan over in $|\hat{p} - p| \leq z\sqrt{(\hat{p}\,\hat{q}/n)}$.
We hebben dus bewezen:

> Als $\hat{p} = S_n/n$, dan ligt p met kans $2\Phi(z) - 1$ in het interval
> $\hat{p} - z\sqrt{(\hat{p}\,\hat{q}/n)} \leq p \leq \hat{p} + z\sqrt{(\hat{p}\,\hat{q}/n)}$.

Dit interval wordt het <u>betrouwbaarheidsinterval</u> genoemd.

<u>Voorbeeld</u>
Om in een bepaald gebied alle inwoners met bloedgroep 0 te bepalen worden 200 personen onderzocht. Resultaat: 80 personen hebben bloedgroep 0. De schatting voor p is dus $\hat{p}$ = 80/200 = 0,40. Wat kunnen we nu zeggen over de werkelijke waarde van p?

Antwoord:
We kiezen z = 2 zodat de kans 0,95 is dat de werkelijke waarde van p tussen 0,40 - $2\sqrt{(0{,}40\text{x}0{,}60/200)} \approx 0{,}33$ en 0,40 + $2\sqrt{(0{,}40\text{x}0{,}60/200)} \approx 0{,}47$ ligt.

<u>Een mooi voorbeeld</u>
Zie het volgende krantenbericht uit 1981:

Nitro-prusside voorkomt doden

In de benadering van harten vaatziekten met behulp van geneesmiddelen zijn twee nieuwe ontwikkelingen te signaleren:

de toepassing van nitroprusside, dat thans in de cardiologische kliniek van het Wilhelmina Gasthuis te Amsterdam aan alle patiënten die een verhoogd risico op een hartinfarct hebben, wordt gegeven;

de calcium-antagonisten, die in enkele soorten in ons land worden toegepast en die spasmen zowel in de hartspier zelf als in de kransslagaderen kunnen temperen.

Nitro-prusside kan als een belangrijke ontwikkeling worden beschouwd, zegt prof. Durrer. Zijn zoon, J.D. Durrer, heeft in de cardiologische kliniek van het Wilhelmina Gasthuis een onderzoek gedaan bij patiënten bij wie kort tevoren een infarct was opgetreden en die bij de Eerste Harthulp in de zgn. 'coronary care' werden opgenomen.

Deze studie die in 1978 aanving, is thans voltooid en uitgewerkt. Onder deze patiënten bij wie de mortaliteit niet gering is, trad een significante lagere sterfte op bij degenen die nitro-prusside kregen toegediend.

De studie omvatte 328 patiënten, van wie 163 nitroprussiede kregen en de andere 165 de zgn. controlgroep niet.

In de eerste groep deden zich in de eerste week na opname 5 sterfgevallen voor, in de controlegroep 18. Conclusie: nitro-prusside helpt ritmestoornissen voorkomen, het verbetert de biochemische hoedanigheden van het hart en de mate van beschadiging bij een eventueel optredend infarct blijft beperkt.

Zodra duidelijk werd dat nitro-prusside zo'n belangrijke invloed op de sterfte had is het dubbelblinde onderzoek (van J.D. Durrer, K.I. Lie en F.J. van Capelle) gestaakt en omgezet is een longitudinale studie, wat wil zeggen dat alle patiënten die met risico op hartinfarct worden opgenomen het middel nu krijgen toegediend en dat hun wederwaardigheden ook op de langere termijn worden gevolgd, zodat ook hierover meer bekend wordt.

Op de calciumantagonisten hopen we in een volgend artikel nader in te gaan.

Op grond van dit bericht is de volgende viervakkentabel op te stellen:

	overleden	niet overleden	
nitro-prusside	**5**	**158**	163
placebo	**18**	**147**	165
	23	305	328

1. Toetsing volgens de exacte test van Fisher (zie par. 7)

Stel H_0 waar, d.w.z. nitro-prusside is niet beter dan een placebo. Dan zullen 305 patiënten overleven ongeacht de behandeling. Onder de 163 patiënten van de proefgroep (de groep die nitro-prusside kreeg) zitten dus 158 van deze 305 patiënten. De kans dat minstens 158 van deze 305 patiënten in de proefgroep zitten is gelijk aan:

$$\alpha = \frac{\binom{305}{158}\binom{23}{5}+\binom{305}{159}\binom{23}{4}+\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots+\binom{305}{163}\binom{23}{0}}{\binom{328}{163}}$$

Deze berekening is desgewenst nog wel uitvoerbaar, maar wordt hier achterwege gelaten.

2. M.b.v de chi-2 toets (zie par. 10)

Als de nulhypothese waar is, dan wordt de verwachte tabel

	overleden	niet overleden
nitro-prusside	11	152
placebo	12	153

De verwachte aantallen zijn minstens 5, dus er is voldaan aan de voorwaarden voor toepassing van de chi-2 toets, zij het niet ruim.

Berekening geeft dan $\chi^2 = 6{,}74$. Met f=1 geeft de tabel $P(\chi^2 \geq 6{,}63) = 0{,}01$, zodat de significantie duidelijk is. De nulhypothese mag dus verworpen worden. Nitro-prusside helpt.

3. Vraag (zie par. 12.1)

Aan de lezer nu de vraag of de volgende redenering ook is toegestaan?
Een schatting van de sterftekans met nitro-prusside is $\hat{p} = \frac{5}{163}$ = 0,031 en met placebo $\hat{p} = \frac{18}{165}$ = 0,109.

De werkelijke sterftekans met nitro-prusside is p = $\hat{p} \pm 2\sqrt{\frac{\hat{p}\,\hat{q}}{n}}$ = 0,031 $\pm$ $2\sqrt{\frac{0{,}031\ x\ 0{,}069}{163}}$

= 0,031 $\pm$ 0,007. De kans hierop is 95%.
Met placebo wordt dit daarentegen p = 0,109 $\pm$ 0,049, ook met kans 95% hierop.
De sterftekans met nitro-prusside ligt dus tussen 0,024 en 0,038.
Met placebo daarentegen tussen 0,06 en 0,158. Dit interval ligt verder naar rechts en geheel buiten het eerste interval. Dus nitro-prusside is overtuigend beter.

12.2. Bepaling van de steekproefgrootte

Voorbeeld
De GGD van een bepaalde stad wil de fractie p van inwoners met luchtwegaandoeningen weten, met ten hoogste een fout van 0,02 in p.
Hoe groot moet de steekproef zijn?

Oplossing:
We kiezen uiteraard z = 2 en dan moet de steekproefgrootte n voldoen aan $2\sqrt{(pq/n)} \leq 0{,}02$.
Helaas, p (en dus ook q) zijn nog niet bekend!
De GGD begint alvast met een steekproef van 500. Dan blijkt: 98 personen hebben luchtwegaandoeningen. Dus $\hat{p}$ = 98/500 ≈ 0,20. Dan is het voldoende om te eisen:
$2\sqrt{(\hat{p}\,\hat{q}/n)} \leq 0{,}02$. Dit betekent n ≥ 1600.
Na deze steekproef blijkt: 332 personen hebben luchtwegaandoeningen.
Conclusie: p = 332/1600 ± 0,02 = 0,21 ± 0,02.

Zonder deze aanpak hadden we moeten eisen: $2\sqrt{(pq/n)} \leq 0{,}02$ ofwel $n \geq 10^4\, p(1-p)$ en dit zou betekenen $n \geq 10^4 \times ¼ = 2500$, omdat p(1 - p) maximaal ¼ kan zijn.

Dus: steekproeven kunnen vaak in omvang worden verkleind door de steekproef in deze twee stappen uit te voeren.

12.3. Zwakke en sterke wet van de grote aantallen

Frequentiedefinitie van kans

$P(A) = \lim_{n\to\infty} \frac{A_n}{n}$ waarbij A_n het aantal keren A bij n herhalingen van het experiment is.

(von Mises, 1920)
Bestaat deze limiet echter wel?
Voor de binomiale verdeling betekent dit $\lim_{n\to\infty} \frac{S_n}{n} = p$.
Geldt dit?
Wel, $E(S_n) = \mu = np$ en $V(S_n) = \sigma^2 = npq$
$\left| \frac{S_n}{n} - p \right| < \varepsilon$ betekent $|S_n - np| < \varepsilon\, n = \frac{\varepsilon\sqrt{n}}{\sqrt{(pq)}}\, \sigma = z\, \sigma$

en de kans hierop is bij benadering $2\,\Phi(z) - 1$.
Voor $n\to\infty$ wordt z zeer groot en nadert deze kans naar $2\,\Phi(\infty) - 1 = 1$.
Dus bewezen:

> Voor willekeurige $\epsilon > 0$ geldt $\lim_{n\to\infty} P(|\frac{S_n}{n} - p| < \varepsilon) = 1$.

Zwakke wet van de grote aantallen (J. Bernoulli, 1713)
Op deze wet berust de mogelijkheid kansen te benaderen via relatieve frequenties.
De stabiliteit van de relatieve frequentie voor veel herhalingen is bewezen.

De kans dat de relatieve frequentie $\frac{S_n}{n}$ minder dan ϵ van p verschilt streeft voor n → ∞ naar 1. Afwijkingen van enig belang tussen $\frac{S_n}{n}$ en p worden voor grote n zeer onwaarschijnlijk.

Sterke wet van de grote aantallen (E. Borel, 1909)

$P(\lim_{n\to\infty} \frac{S_n}{n} = p) = 1$

Geen bewijs.
Deze wet is sterker dan de zwakke wet van van de grote aantallen.

Zij X een willekeurige toevalsvariabele op Ω.
en zij A = {ω ϵ Ω | |X(ω)| ≥ ϵ }.
Dan $E(I_A) = P(A) = P(|X(\omega)| \geq \epsilon)$.
Er geldt $|X(\omega)| \geq \epsilon\ I_A(\omega)$, dus ook $X(\omega)^2 \geq \epsilon^2\ I_A(\omega)$ en dus ook $E(X^2) \geq \epsilon^2\ E(I_A) = \epsilon^2\ P(A)$
zodat $P(A) \leq \epsilon^{-2}\ E(X^2)$ voor elke ϵ > 0 ofwel:

$P(\|X\| \geq \epsilon) \leq \epsilon^{-2}\ E(X^2)$ voor elke ϵ > 0

Ongelijkheid van Chebychev (1821 – 1894)

$X \to X - \mu$: $P(|X-\mu| \geq \epsilon) \leq \frac{E((X-\mu)2)}{\varepsilon^2} = \frac{V(X)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}$

$\varepsilon \to t\sigma$:

$P(\|X-\mu\| \geq t\sigma) \leq \frac{1}{t^2}$ voor elke t > 0

Een opvallend en onverwacht resultaat! De kans dat een toevalsvariabele meer dan tσ afwijkt van μ is kleiner dan $1/t^2$.
Hier zie je dus wat het grote belang van σ is.

13. Hypergeometrische verdeling

Uit N knikkers, waarvan r zwarte en N - r witte, nemen we een steekproef van n.
Dit is trekking zonder teruglegging. Bij de binomiale verdeling ging het om trekking met teruglegging.
Zij S_n het aantal zwarte knikkers in de steekproef.
Dan geldt:

$$P(S_n = x) = h(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \qquad x = 0, 1, \ldots, r$$

$\mu = E(S_n) = np$ met $p = r/N$
$\sigma^2 = V(S_n) = npq\,\frac{N-n}{N-1}$

Bewijs:

Stel de n zwarte knikkers zijn genummerd: 1 t/m n.
Def. $X_i = 1$ als nr. i in de steekpr. valt, anders $= 0$.
Dan $S_n = X_1 + \cdots + X_n$

$$E(X_i) = P(\text{nr. } i \text{ in de steekpr.}) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}, \text{ dus } E(S_n) = n \cdot \frac{n}{N} = np.$$

$$S_n^2 = (X_1 + \cdots + X_n)^2 = \sum_i X_i^2 + 2\sum_{i<j} X_i X_j = \sum_i X_i + 2\sum_{i<j} X_i X_j.$$

$$E(X_i X_j) = P(X_i X_j = 1) = P(X_i = 1 \wedge X_j = 1) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}, \text{ dus}$$

$$E(S_n^2) - E(S_n)^2 = np + 2\binom{n}{2}\frac{n(n-1)}{N(N-1)} - (np)^2 = np\left(1 + \frac{(n-1)(n-1)}{N-1} - \frac{n \cdot n}{N}\right) = np\,\frac{N(N-1) + (n-1)(n-1)N - nn(N-1)}{N(N-1)}$$

$$= np\,\frac{N^2 - N + (nN - n - n + 1)N - nn(N-1)}{N(N-1)} = np\,\frac{N^2 - nN - mN + nm}{N(N-1)} = np\,\frac{(N-m)(N-n)}{N(N-1)} = npq\,\frac{N-m}{N-1}.$$

Voorbeeld
In een zending van 100 exemplaren zitten 80 goede en 20 defecte exemplaren.
De ontvanger weet dat niet en neemt een steekproef van 10 exemplaren. Resultaat: 2 defecte exemplaren.
Hoe groot is de kans op 2 defecte exemplaren?

Antwoord:

$N = 100,\ r = 20,\ n = 10$ en $x = 2$

Dan $h(2) = \dfrac{\binom{20}{2}\binom{80}{8}}{\binom{100}{10}} \approx 0{,}318$ volgens tabel, met 3 parameters N, r en n.

Dit voorbeeld is echter niet echt realistisch! De ontvanger wil graag een uitspraak doen wat nu de fractie defecte exemplaren kan zijn.

Een interessante voorbeeld

Bij veel praktische vragen is N onbekend en wil men N schatten op grond van een steekproef.

Stel dat men het aantal vissen in een vijver wil schatten. Laat N het onbekende aantal vissen zijn. Men vangt r vissen, merkt deze en zet ze daarna terug. Na enkele dagen wordt een steekproef van n vissen genomen. Daar blijken x gemerkte vissen bij te zijn.

Is nu een schatting van N te geven?

Wel, de verhouding x : n is een benadering van de werkelijke verhouding r : N.

Dus x/n is een schatting van r/N, zodat r n / x een schatting van N is. Hoe goed is deze schatting nu?

Uit de definitie van h(x) volgt $\dfrac{h(x,N)}{h(x,N-1)} = \dfrac{(N-r)(N-n)}{(N-r-n+x)N} \geq 1$ als $N \leq \dfrac{r\,n}{x}$

< 1 als $N > \dfrac{r\,n}{x}$

zodat h maximaal is als $N = \dfrac{r\,n}{x}$.

Stelling

Als N veel groter is dan n (vuistregel N > 10 n), dan maakt het niet veel uit of men teruglegt of niet en kan men i.p.v. h(x) de binomiaalverdeling b(x) nemen.

Dus:

$h(x) \approx b(x) = P(S_n = x) = \binom{n}{x}\ (r/N)^x\ (1 - r/N)^{n-x}$ met $\mu = np$ en $\sigma^2 \approx npq$ (omdat $\dfrac{N-n}{N-1} \approx 1$).

14. Pascal-verdeling (negatieve binomiale verdeling)

Experiment met succeskans p herhalen tot n successen zijn opgetreden.
T_n: aantal opgetreden missers.
Kansverdeling?

Als het aantal opgetreden missers x is, dan zijn er x+n herhalingen geweest, waarbij er x missers en n-1 successen waren in de eerste x+n-1 herhalingen en daarna nog een succes.
De kans hierop is $\binom{x+n-1}{x} q^x p^n$ voor x = 0, 1, 2, ….. Dus:

$P(T_n = x) = \binom{n-1+x}{x} p^n q^x$ voor x = 0, 1, 2, …..

Pascal-verdeling

Deze formule is ook zo te schrijven (ga na):

$P(T_n = x) = (-1)^x \binom{-n}{x} p^n q^x$

Vandaar de naam negatieve binomiaalverdeling.
De som van alle kansen bij de Pascal-verdeling is inderdaad 1, zoals men kan uitrekenen.

Stelling

$E(T_n) = np^{-1}q$ en $V(T_n) = np^{-2}q$

Bewijs:

$$E(T_n) = \sum_{x=0}^{\infty} x P(T_n = x) = \sum_{x=1}^{\infty} x (-1)^x \binom{-n}{x} p^n q^x = p^n \sum_{x=1}^{\infty} \frac{(-n)!}{(x-1)!(-n-x)!} (-q)^x \qquad y = x-1$$

$$= p^n \sum_{y=0}^{\infty} \frac{(-n)!}{y!(-n-1-y)!} (-q)^{y+1} = -qp^n \sum_{y=0}^{\infty} (-n) \binom{-n-1}{y} (-q)^y = nqp^n (1-q)^{-n-1}$$

$$= nqp^n p^{-n-1} = \boxed{np^{-1}q}$$

De afleiding van de variantie gaat gemakkelijker als we gebruik maken van momenten.
Zie hiervoor de betreffende par.

Speciaal geval van de Pascal-verdeling: n =1.
Dan krijgen we de geometrische verdeling:

$P(T_1 = x) = pq^x$ voor x = 0,1,2, ……..

met $E(T_1) = p^{-1} q$ en $V(T_1) = p^{-2} q$.

Opmerking
De toevalsvariabele T_n wordt ook wel iets anders gedefiëerd:
Zij X het aantal worpen dat nodig is om n successen te bereiken. Dan geldt $X = T_n + n$.
Dan $E(X) = E(T_n) + n = np^{-1}q + n = \frac{n}{p}$ en $V(X) = V(T_n) = np^{-2}q$.

Voorbeeld van de Pascal-verdeling
vdBrink/Koele blz. 75
Van een bestaande operatie is de kans op succes 0,5. Van een nieuwe techiek wordt beweerd dat de kans op suces 0,8 is. Men besluit deze claim te onderzoek. In verband met de kosten en de risico's liefst met een zo klein mogelijk aantal patiënten. Men besluit volgens de nieuwe techniek te opereren tot het 5^e succes is behaald en dan een conclusie te trekken.
Als p = 0,5 dan is het verwachte aantal operaties dat nodig is om 5 successen te behalen gelijk aan 5/0,5 = 10. Als p = 0,6 dan is het verwachte aantal gelijk aan 5/0,8 ≈ 6.
Resultaat van het onderzoek: de 10^e operatie levert het 5^e succes op. Dit resultaat pleit veel meer voor p = 0,5 dan voor p = 0,6. Het lijkt dus verstandig de bestaande operatie te handhaven.

15. Meer kansverdelingen

In de afgelopen decennia zijn veel meer kansverdelingen bestudeerd die in de statistiek gebruikt kunnen worden.
We noemen hier slechts een klein aantal bekende en veelgebruikte kansverdelingen.

t-verdeling (Student-verdeling)
Voor kleinere steekproeven waarbij de variatie onbekend is.

F-verdeling
Gebruik vooral in de variantie-analyse.

Cauchy-verdeling
Bijzonder geval van een t-verdeling, waarbij μ en σ onbekend zijn.

Gamma-verdeling
Wordt veel gebruikt bij wachtrij-analyses.

Deze kansverdelingen zijn zeer belangrijk in de statistiek. Statistiek is een discipline die gaat over het schatten van kansen uit waarnemingen. Een discipline die een enorme ontwikkeling doorgemaakt heeft en waar binnen de wiskunde al lijvige studieboeken over bestaan. Het vak statistiek is onderhand ook een belangrijke hulpwetenschap geworden in de psychologie, pedagogiek, biologie, geneeskunde , enz., omdat daar natuurlijk heel vaak experimenten worden uitgevoerd die data opleveren waar men conclusies uit wil kunnen trekken. De daarbij gebruikte statistische methoden variëren van eenvoudig tot geavanceerd niveau. Omdat de achterliggende wiskundige achtergronden in het algemeen ver buiten het bereik van zowel studenten als wetenschappelijke staf in de onderhavige vakgebieden liggen, zit er niets anders op dan dat de veel gebruikte statistische methoden (op het vlak van toetsen, schatten, steekproefverdelingen, enz.) slechts gepresenteerd kunnen worden in de vorm van “kookboekrecepten”. Het gevolg is dat het vak statistiek vaak een lastig bijvak is in de menswetenschappen.
Statistiek wordt binnen de studie wiskunde vaak als toegepaste wiskunde beschouwd, mede doordat statistiek een wetenschap is waar het begrip onzekerheid een dominante rol speelt.

Kansrekening daarentegen gaat over het berekenen van kansen en dat spreekt wiskundestudenten vaak meer aan. Het mag dan over kansen gaan, maar deze worden wel exact berekend.

Het doel van deze syllabus is om de belangrijkste begrippen uit de kansrekening te behandelen en alleen een inleiding in de statistiek te geven. Geavanceerde technieken uit de statistiek zullen hier dus niet aan de orde komen.

16. Toevalsexperimenten met een aftelbaar oneindig aantal uitkomsten

Voorbeeld
Werpen met een zuivere dobbelsteen tot een zes valt.
Gebeurt dit wel ooit?
Mogelijke uitkomsten: 6, ≠6 6, ≠6 ≠6 6, ……
Kansen resp.: 1/6, 5/6x1/6, $(5/6)^2$x1/6, ……..
Er zijn dus aftelbaar oneindig veel mogelijke uitkomsten.
De som van alle kansen is 1/6 + 5/6x1/6 + $(5/6)^2$x1/6 + ……. = $\frac{1/6}{1-5/6}$ = 1.
Dus aan deze eis voor een kansfunctie is voldaan.
Het blijkt dat we de bekende regels uit de kansrekening (voor toevalsexperimenten met een eindig aantal mogelijke uitkomsten) gewoon kunnen toepassen. De theoretische achtergronden zullen we hier niet verder behandelen.

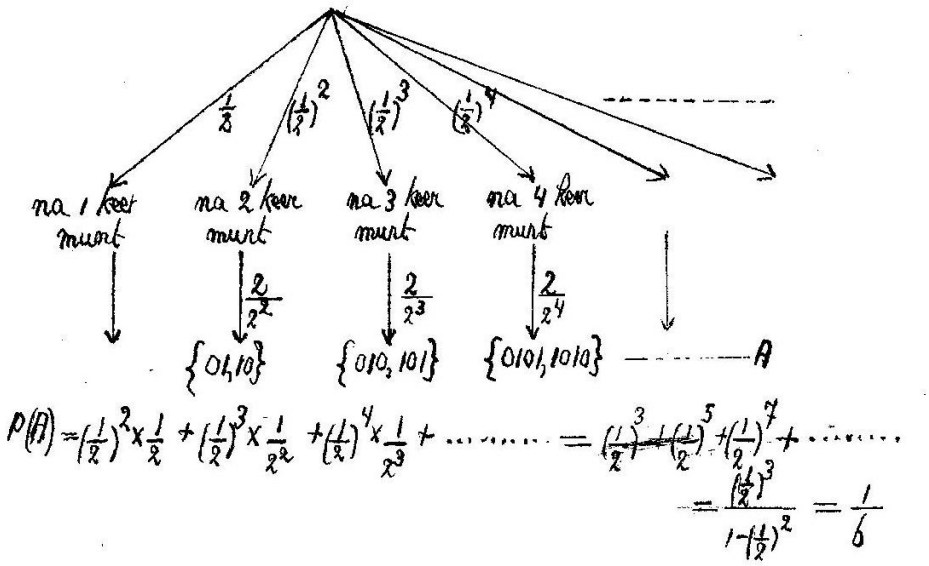
Voorbeeld
Aan een geluksrad, met p de kans op 1, wordt net zo vaak gedraaid totdat de 1 wordt bereikt. Zij N het benodigde aantal draaiingen. Bereken E(N).

Antwoord:
N=1: gunstig is de uitkomst 1. Kans: p.
N=2: gunstig is de uitkomst 0,1. Kans: (1-p)p.
N=3: gunstig is de uitkomst 0,0,1. Kans: $(1-p)^2$p.
Enz.
Dus E(N) = 1.p + 2.(1-p)p + 3.$(1-p)^2$p + ……. = p.S
S = 1 + 2.(1-p) + 3.$(1-p)^2$ + ……
(1-p)S = 1-p + 2.$(1-p)^2$ + 3.$(1-p)^3$ + …….
Aftrekken levert: pS = 1 + (1-p) + $(1-p)^2$ + …… = $\frac{1}{1-(1-p)}$ = $\frac{1}{p}$
Derhalve E(N) = pS = $\frac{1}{p}$
Een verrassend resultaat! Het gaat hier om een Pascal-verdeling met n = 1 (zie 2e definitie van de Pascal-verdeling).

Voorbeeld
Werpen met een zuivere munt tot munt valt.
Mogelijke uitkomsten: m, km, kkm, ……
Kansen resp.: ½, $(½)^2$, $(½)^3$, ……..
Ook nu geldt ∑ p(ω = $\frac{½}{1-½}$) = 1.
Is het mogelijk dat iemand nooit munt gooit?
Dit is fysisch niet waar te nemen.
De kans dat n worpen nodig zijn is $(½)^n$. Voor n→∞ is de limiet 0.
Conclusie?
Zij X het aantal worpen dat nodig is totdat munt valt.
Dan E(X) = ∑ xp(x) = 1 x p(1) + 2 x p(2) + 3 x p(3) + …… = 1x½ + 2x$(½)^2$ + 3x$(½)^3$ + ……
= 2 (ga na m.b.v. meetkundige reeksen).

Voorbeeld
Werpen met een zuivere munt tot munt valt.
Als dit na n keer is, dan daarna nog n keer werpen met de zuivere munt.
Gevraagd: de kans op de gebeurtenis A dat deze tweede serie van n worpen een alternerende serie te zien geeft.

Antwoord:
Het gereduceerde kansdiagram wordt:



$$P(A) = \left(\tfrac{1}{2}\right)^2 \times \tfrac{1}{2} + \left(\tfrac{1}{2}\right)^3 \times \tfrac{1}{2^2} + \left(\tfrac{1}{2}\right)^4 \times \tfrac{1}{2^3} + \ldots\ldots = \left(\tfrac{1}{2}\right)^3 + \left(\tfrac{1}{2}\right)^5 + \left(\tfrac{1}{2}\right)^7 + \ldots\ldots$$

$$= \frac{\left(\frac{1}{2}\right)^3}{1-\left(\frac{1}{2}\right)^2} = \frac{1}{6}$$

Maar, mogen we zo wel redeneren?

<u>Sint-Petersburg paradox</u>
In een casino kan na betaling het volgende spel worden gespeeld. Het casino legt 1 euro in en daarna werpt de speler met een zuivere munt. Is de uitkomst munt, dan krijgt de speler de inleg en is het spel afgelopen. Is de uitkomst kruis, dan verdubbelt het casino de inleg en mag de speler opnieuw werpen. Bij munt krijgt de speler de pot. Bij kruis verdubbelt het casino opnieuw de inleg. Enz.
De vraag is nu hoeveel de speler bereid is te betalen om aan dit spel mee te doen.
In de praktijk zullen maar weinig mensen bereid zijn voor een groot bedrag mee te doen, omdat men denkt voor dit bedrag niet veel terug te krijgen.
Wat kan de kansrekening ons nu over deze kwestie leren?

Zij X het aantal worpen dat het spel duurt en Y de uitbetaling door het casino.
De bijbehorende kansverdelingen zijn:

x	1	2	3	4	5	
p(x)	1/2	$(1/2)^2$	$(1/2)^3$	$(1/2)^4$	$(1/2)^5$	…….
y	1	2	2^2	2^3	2^4	…….

Dus:
$E(Y) = \sum yp(y) = \sum yp(x) = 1 \times 1/2 + 2 \times (1/2)^2 + 2^2 \times (1/2)^3 + 2^3 \times (1/2)^4 + 2^4 \times (1/2)^5 + \ldots\ldots$
$= 1/2 + 1/2 + 1/2 + \ldots\ldots = \infty$.
De verwachtingswaarde van de uitbetaling is oneindig groot, dus de gemiddelde uitbetaling zal oneindig groot zijn. Dus, het spel is voordelig voor de speler. Dus wel spelen, zou je dan zeggen!
Maar, een uitbetaling van $2^4 = 16$ euro heeft slechts een kans ter grootte van $(1/2)^5 = 1/32$, dus ca. 3%. Hoe groter de uitbetaling, hoe kleiner de kans.
Overigens, het spel duurt maar kort. De verwachtingswaarde van X is:
$E(X) = \sum xp(x) = 1 \times 1/2 + 2 \times (1/2)^2 + 3 \times (1/2)^3 + 4 \times (1/2)^4 + \ldots\ldots = 2$ (zie een vorig voorbeeld).
Nog sterker: als het spel lang genoeg duurt, dan kan het casino de verdubbeling ook niet meer volhouden.

Deze paradox werd in 1738 aan de Academie voor Wetenschappen in Sint-Petersburg voorgelegd door D. Bernoulli.

Voorbeeld
Werpen met een zuivere munt tot 2 keer munt valt.
Zij X het aantal benodigde worpen.
Gevraagd: E(X).

Antwoord:

$E(X) = \sum xp(x) = 2p(2) + 3p(3) + 4p(4) + \ldots\ldots\ldots = 2(1/2)^2 + 3\binom{2}{1}(1/2)^3 + 4\binom{3}{1}(1/2)^4 + 5\binom{4}{1}(1/2)^5 + \ldots\ldots = ?$ Probeer dit zelf eens uit te rekenen.

Volgens Pascal geldt $E(X) = \frac{n}{p} = \frac{2}{½} = 4$.

Voorbeeld
Uit $Z^+ = \{1,2,3,\ldots\ldots\}$ aselect een nummer trekken. Is logisch wel duidelijk, maar fysisch niet
Alle kansen p(ω) zouden gelijk moeten zijn. Maar P(Ω) = 1, dus p(ω) is niet gedefiniëerd?
Maar er zijn wel deelverzamelingen A van Ω waarvoor P(A) wel gedefiniëerd is.
Bijv.

Als A = {2,4,6,…… }, dan $P(A) = \frac{1}{2}$

Als A = {3,6,9,..……}, dan $P(A) = \frac{1}{3}$

Als A = {4,8,12,…....}, dan $P(A) = \frac{1}{4}$

Enz.
Het wordt dus nu al ingewikkelder!
In par. 2.2 komen we hier nog even op terug.

17. Toevalsexperimenten met een overaftelbaar aantal uitkomsten

We herhalen nog even:

Eindig aantal uitkomsten
X: $\Omega \rightarrow R$ toevalsvariabele

Verwachtingswaarde van X: $E(X) = \sum_{i=1}^{n} x_i\, p(x_i) = \sum_{i=1}^{n} x_i\, P(X=x_i)$

Aftelbaar oneindig veel uitomsten
Een toevalsvariabele X: $\Omega \rightarrow R$ kan dan eindig veel of aftelbaar oneindig veel waarden hebben.

Verwachtingswaarde van X: $E(X) = \sum_{i=1}^{n} x_i\, p(x_i) = \sum_{i=1}^{n} x_i\, P(X=x_i)$

Of

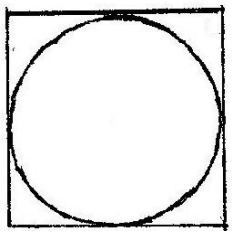
Verwachtingswaarde van X: $E(X) = \sum_{i=1}^{\infty} x_i\, p(x_i) = \sum_{i=1}^{\infty} x_i\, P(X=x_i)$

Voorwaarde: de reeks moet absoluut convergent zijn.

Overaftelbaar veel uitkomsten
Dan heeft de definitie $P(A) = \sum_{\omega \epsilon A} p(\omega)$ geen betekenis, omdat zulke sommeringen niet kunnen.
Toch zijn er genoeg toevalsexperimenten met overaftelbare uitkomstenverzameling waar we P(A) kunnen definiëren voor zekere $A \subset \Omega$.

Voorbeeld: oppoten
Voetlengte a tegen voetlengte b. Welke winstkansen horen bij a en b?
Als a begint, dan ontstaat het patroon a,b,a,b,a,…..
Op een gegeven moment zal een stuk x resten waar a en b niet beide meer in passen. Dus x<a+b.
Als a aan beurt is, dan is gunstig voor a: $a \leq x < a+b$. En gunstig voor b: $x < a$.
Dan geldt: P(a wint) = lengte [a,a+b) / lengte [0,a+b) = b/(a+b).
Of de eindpunten wel of niet meetellen is niet relevant.
Dus in dat geval: P(b wint) = 1 – b/(a+b) = a/(a+b).
Analoog: als b aan beurt is, dan geldt P(b wint) = a/(a+b) en P(a wint) = b/(a+b).
Derhalve: P(a wint) = b/(a+b) en P(b wint) = a/(a+b), onafhankelijk van wie er begint.
Conclusie: de kleinste voetlengte gaat winnen, onafhankelijk van wie er begint.

Voorbeeld



Cirkel met straal r.
Kies aselect een punt in het vierkant.
Dan P(punt valt binnen de cirkel) = opp. cirkel / opp. vierkant = $\pi r^2 / (2r)^2 = \frac{1}{4}\pi$.

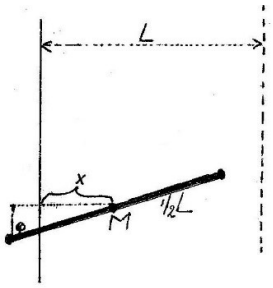
N.B.
In het algemeen geldt dat we P(A) ongedefiniëerd moeten laten voor zekere A $\subset$ Ω en dat we ons moeten beperken tot een zekere collectie van deelverzamelingen van Ω.

Voorbeeld
Experiment: kies het punt P aselect op het interval [a,b].
Dan Ω = [a,b], dus Ω is op te vatten als de verzameling waarden op een gesloten interval.
Als A een willekeurig deelinterval [x,y] van [a,b] is, dan is de kans op A ongetwijfeld te definiëren als P(A) = $\frac{y-x}{b-a}$.

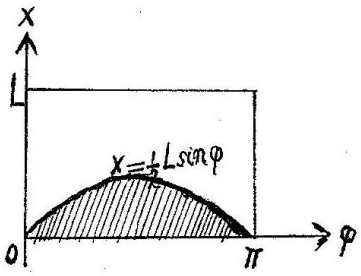
Voorbeeld
Experiment: kies punt P aselect op het cirkeloppervlak van een cirkel met middelpunt M en straal r.
Als A nu een deel van dit cirkeloppervlak is en aan A is een oppervlakte toe te kennen, dan zou de kans op A wellicht gedefiniëerd kunnen worden als P(A) = $\frac{opp.\ A}{\pi r^2}$.
Als we zekerheid willen hebben, dan moeten we meer weten wanneer een kansdefinitie zinvol en consistent is. In par. 22 zullen we daar nader op ingaan.

Naald van Buffon
Op een vlak zijn evenwijdige stroken aangebracht met strookbreedte 2L. Men laat willekeurig een naald met lengte L vallen op het vlak en kijkt dan of de naald een der evenwijdige lijnen snijdt.
Gevraagd: de kans dat er snijding optreedt.

Oplossing:
We kunnen het experiment vereenvoudigen tot het volgende:
We nemen slechts één strook met breedte 2L en laten de naald dan willekeurig zo vallen dat het midden M van de naald op de linkerhelft van de strook valt. De kans dat de naald dan de linkerlijn van de strook snijdt is gelijk aan de gevraagde kans.
Zij x de afstand van M tot de linkerlijn en φ de hoek die de naald maakt met de linkerlijn.
Dan is mogelijk: x ≤ L en 0 ≤ φ ≤ π.
Gunstig voor snijding van de naald met de linkerlijn is: x ≤ ½ L sin φ.



Hieronder het mogelijke gebied versus het gunstige (= gearceerde) gebied:



Hieruit volgt:

$$P(\text{snijding}) = \text{opp. gearc. gebied / opp. rechthoek} = \int_0^{\pi} \tfrac{1}{2} L \sin\varphi \, d\varphi \; / \; \pi L = \frac{1}{\pi}$$

Herkomst:
Dit probleem is afkomstig van Georg-Louis Leclerc, graaf van Buffon, 1777.
Verschillende wiskundigen hebben in de jaren daarna aandacht besteed aan dit probleem.
Dit is te begrijpen als men bedenkt dat:

- de kansrekening destijds nog in de kinderschoenen stond;
- er van oudsher altijd al fascinatie voor de grootheid pi bestond.

Meerdere keren zijn naaldexperimenten uitgevoerd om hiermee een aantal decimalen van pi te vinden. O.a. in 1901 door de Italiaanse wiskundeleraar Lazzarini. Zijn resultaten waren echter zo mooi dat ze de aandacht trokken. Uiteindelijk kon overtuigend worden bewezen dat hier sprake was van bedachte resultaten. Ook toen kwam fraude in de wetenschap al voor.

Hoe wordt nu bij een toevalsexperiment met overaftelbaar veel mogelijke uitkomsten het begrip kans ingevoerd?
Antwoord:
Niet p(a), maar P(A) wordt gedefiniëerd voor bepaalde deelverzamelingen A, die we dan gebeurtenissen noemen.
In de laatste par. wordt hier nog wat verder op ingegaan.

18. Momenten

Zij X een toevalsvariabele.
Onder het n^e moment van X wordt verstaan:

$E(X^n) = \sum x^n\, p(x)$ waarbij x alle waarden van X doorloopt.

Als de uitkomstenverzameling eindig is, dan zijn er ook slechts eindig veel waarden van X.
Nauwkeuriger formulering:

$E(X^n) = \sum x_i^n\, p(x_i) = \sum x_i^n\, P(X=x_i)$.

Onder de momentvoortbrengende functie van X wordt verstaan:

$M(\vartheta) = E(e^{\vartheta X}) = \sum e^{\vartheta x_i} p(x_i)$

Niet elke toevalsvariabele heeft een momentvoortbrengende functie!

Stelling
$M(\vartheta) = 1 + \vartheta\, E(X) + \vartheta^2/2!\, E(X^2) + \ldots\ldots..$

Geen bewijs.
De reeks is convergent.

Stelling
Voor de n^e afgeleide van het moment geldt:

$M^{(n)}(0) = E(X^n)$

Geen bewijs.
Wel als volgt in te zien: differentiëer in de vorige stelling n keer onder het $\sum$-teken.

Deze stelling maakt het veel gemakkelijker om $E(X^2)$, $E(X^3)$, enz. te berekenen.
Dus ook V(X).

Binomiale verdeling
$P(S_n=x) = b(x) = \binom{n}{x} p^x q^{n-x}$ voor x = 0,1,2,……., met $\mu = E(S_n) = np$ en $V = \sigma^2 = npq$
Nu de berekening m.b.v. momenten:

$$M(\vartheta) = E(e^{\vartheta S_n}) = \sum_{x=0}^{n} e^{\vartheta x} b(x) = \sum_{x=0}^{n} e^{\vartheta x} \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (pe^{\vartheta})^x q^{n-x} = (pe^{\vartheta}+q)^n$$

$$M'(\vartheta) = n(pe^{\vartheta}+q)^{n-1} \cdot pe^{\vartheta} \quad \text{dus} \quad \mu = M'(0) = n(p+q)^{n-1} \cdot p = \boxed{np}$$

$$M''(\vartheta) = n(n-1)(pe^{\vartheta}+q)^{n-2}(pe^{\vartheta})^2 + n(pe^{\vartheta}+q)^{n-1} \cdot pe^{\vartheta}$$

$$\text{dus } M''(0) = n(n-1)p^2 + np = n^2p^2 - np^2 + np = n^2p^2 + npq = E(S_n^2)$$

$$\text{zodat } \sigma^2 = E(X^2) - \mu^2 = n^2p^2 + npq - (np)^2 = \boxed{npq}$$

<u>Poisson-verdeling</u>

$P(X = x) = b(x) \approx \frac{\mu^x}{x!} e^{-\mu}$ met x = 0,1,2,3,….. en $X = S_n$, $\mu = E(S_n) = np$ en $\sigma^2 = V(S_n) \approx \mu$.

Berekening m.b.v. momenten:

$$M(\vartheta) = E(e^{\vartheta X}) = \sum_{x=0}^{\infty} e^{\vartheta x} \cdot e^{-\mu} \frac{\mu^x}{x!} = \sum_{x=0}^{\infty} e^{-\mu} \frac{(\mu e^{\vartheta})^x}{x!} = e^{-\mu} e^{\mu e^{\vartheta}} = e^{\mu(e^{\vartheta}-1)}$$

$$M'(\vartheta) = e^{\mu(e^{\vartheta}-1)} \cdot \mu e^{\vartheta} \quad \text{dus} \quad E(X) = M'(0) = e^0 \cdot \mu e^0 = \boxed{\mu}$$

$$M''(\vartheta) = e^{\mu(e^{\vartheta}-1)}(\mu e^{\vartheta})^2 + e^{\mu(e^{\vartheta}-1)} \cdot \mu e^{\vartheta} \quad \text{dus} \quad E(X^2) = M''(0) = \boxed{\mu^2+\mu}$$

$$\sigma^2 = (\mu^2+\mu) - \mu^2 = \boxed{\mu}$$

<u>Pascalverdeling</u>

$P(T_n = x) = \binom{n-1+x}{x} p^n q^x$ voor x = 0,1,2,…….

Alternatieve formulering: $P(T_n = x) = (-1)^x \binom{-n}{x} p^n q^x$

met $E(T_n) = np^{-1}q$ en $V(T_n) = np^{-2}q$

Berekening m.b.v. momenten:

$$M(\vartheta) = E(e^{\vartheta T_n}) = \sum_{x=0}^{\infty} e^{\vartheta x} (-1)^x \binom{-n}{x} p^n q^x = p^n \sum_{x=0}^{\infty} \binom{-n}{x} (-qe^{\vartheta})^x = p^n (1-qe^{\vartheta})^{-n} \quad \text{wenn } qe^{\vartheta} < 1$$

(reeksontwikkeling)

$$M'(\vartheta) = p^n \cdot -n(1-qe^{\vartheta})^{-n-1} \cdot -qe^{\vartheta} = np^n q (1-qe^{\vartheta})^{-n-1} e^{\vartheta}$$

dus $\mu = E(T_n) = M'(0) = np^n q p^{-n-1} = \boxed{np^{-1}q}$

$$M''(\vartheta) = np^n q \left\{ (-n-1)(1-qe^{\vartheta})^{-n-2} \cdot -qe^{\vartheta} \cdot e^{\vartheta} + (1-qe^{\vartheta})^{-n-1} e^{\vartheta} \right\} = np^n q (1-qe^{\vartheta})^{-n-2} e^{\vartheta} (nqe^{\vartheta} + 1)$$

dus $E(T_n^2) = M''(0) = np^n q p^{-n-2} (nq+1) = \boxed{np^{-2}q(nq+1)}$

Derhalwe $\sigma^2 = np^{-2}q(nq+1) - (np^{-1}q)^2 = \boxed{np^{-2}q}$

19. Verdelingsfuncties

Gegeven: een willekeurig toevalsexperiment en X: Ω → R een toevalsvariabele.
Onder de verdelingsfunctie van X wordt de volgende functie verstaan:

$$F(x) = P(X \leq x) = P\{\omega \in \Omega \mid X(\omega) \leq x\}$$

***Mogelijkheid 1**: X heeft hoogstens aftelbaar oneindig veel waarden.*

Dan $F(x) = P(X \leq x) = \sum_{x_i \leq x} P(x = x_i)$.
En $E(X) = \sum_{x_i \leq x} x_i P(x = x_i)$.

Voorbeeld
Ω eindig.
Experiment: worp met 2 zuivere dobbelstenen.
Toevalsvariabele X: het gegooide aantal zessen.
Voor de kansverdeling en verdelingsfunctrie van X geldt:

Kansverdeling van X:

x	p(x)
0	$\frac{25}{36}$
1	$\frac{10}{36}$
2	$\frac{1}{36}$

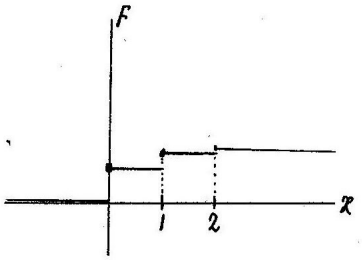
$F(x) = P(X < x) = 0$ voor $x < 0$
$= \frac{25}{36}$ voor $0 < x < 1$
$= \frac{35}{36}$ voor $1 < x < 2$
$= 1$ voor $2 < x$

dus X is discreet (verdeeld), met eindig veel waarden

$\mu = E(X) = 0.\frac{25}{36} + 1.\frac{10}{36} + 2.\frac{1}{36} = \boxed{\frac{1}{3}}$

$E(X^2) = 1^2.\frac{10}{36} + 2^2.\frac{1}{36} = \frac{14}{36}$, dus $V(X) = \frac{14}{36} - \left(\frac{1}{3}\right)^2 = \frac{10}{36}$

zodat $\sigma = \boxed{\frac{1}{6}\sqrt{10}}$



F is niet continu

Voorbeeld
Ω aftelbaar oneindig.
Experiment: werpen met een zuivere munt tot “kruis” is gevallen.
Dus Ω = {1, 01, 001, 0001,…….}.
Zij de toevalsvariabele X het benodigde aantal worpen.
Dan $P(X = x) = (½)^x$, x = 1, 2, 3, ……
Dus X is discreet verdeeld, met aftelbaar oneindig veel waarden.

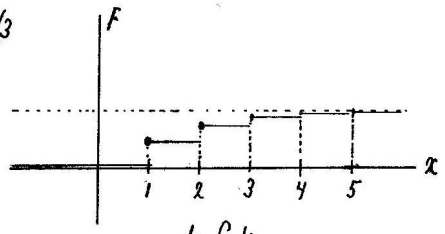
De verdelingsfunctie is de volgende:

$F(x) = P(X \leq x) = 0$ voor $x < 1$

$= \frac{1}{2}$ voor $1 \leq x < 2$

$= \frac{1}{2} + (\frac{1}{2})^2$ voor $2 \leq x < 3$

⋮



trapfunctie

$M(\vartheta) = E(e^{\vartheta X}) = \sum_{x=1}^{\infty} e^{\vartheta x} P(X=x) = \sum_{x=1}^{\infty} e^{\vartheta x} (\frac{1}{2})^x = \sum_{x=1}^{\infty} (\frac{1}{2}e^{\vartheta})^x = \frac{\frac{1}{2}e^{\vartheta}}{1-\frac{1}{2}e^{\vartheta}} = \frac{1}{2e^{-\vartheta}-1} = (2e^{-\vartheta}-1)^{-1}$

$M'(\vartheta) = -(2e^{-\vartheta}-1)^{-2} \cdot -2e^{-\vartheta} = 2e^{-\vartheta}(2e^{-\vartheta}-1)^{-2}$, dus $E(X) = M'(0) = 2(2-1)^{-2} = \boxed{2}$

$M''(\vartheta) = -2e^{-\vartheta}(2e^{-\vartheta}-1)^{-2} + 2e^{-\vartheta} \cdot -2(2e^{-\vartheta}-1)^{-3} \cdot -2e^{-\vartheta}$, dus $E(X^2) = M''(0) = -2+8 = \boxed{6}$

Derhalve $\sigma^2 = 6 - 2^2 = \boxed{2}$

Voorbeeld

Ω is aftelbaar oneindig.

Experiment: werpen met een zuivere munt tot 2 keer “kruis” is gevallen.

Dus Ω = {11, 011, 101, …….}.

Zij X het benodigde aantal worpen.

Dan geldt P(X = x) = $\frac{x-1}{2^x}$, x = 2,3,4, …..

Dus X is discreet verdeeld, met aftelbaar oneindig veel waarden.

De verdelingsfunctie van X is:

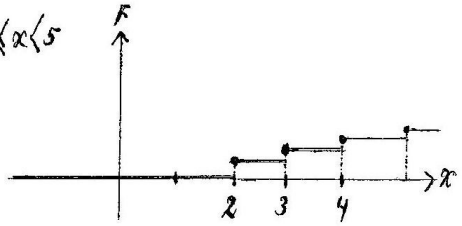
$F(x) = P(X < x) = 0$ als $x < 2$

$= \frac{1}{2^2} = \frac{1}{4}$ als $2 \leq x < 3$

$= \frac{1}{2^2} + \frac{2}{2^3} = \frac{1}{2}$ als $3 \leq x < 4$

$= \frac{1}{2^2} + \frac{2}{2^3} + \frac{3}{2^4} = \frac{11}{16}$ als $4 \leq x < 5$

⋮



Momenten:

$$M(\vartheta) = E\left(e^{\vartheta X}\right) = \sum_{x=2}^{\infty} e^{\vartheta x} P(X=x) = \sum_{x=2}^{\infty} e^{\vartheta x} \frac{x-1}{2^x} = \sum_{x=2}^{\infty} (x-1)a^x \text{ met } a = \frac{1}{2}e^{\vartheta}$$

$$= a^2 + 2a^3 + 3a^4 + 4a^5 + \cdots$$

$$= a^2 + a^3 + a^4 + a^5 + \cdots$$
$$+ a^3 + a^4 + a^5 + \cdots$$
$$+ a^4 + a^5 + \cdots$$
$$+ a^5 + \cdots$$

$$= \frac{a^2}{1-a} + \frac{a^3}{1-a} + \frac{a^4}{1-a} + \frac{a^5}{1-a} + \cdots = \frac{a^2 + a^3 + a^4 + a^5 + \cdots}{1-a} = \frac{\frac{a^2}{1-a}}{1-a} = \left(\frac{a}{1-a}\right)^2 = \left(\frac{\frac{1}{2}e^{\vartheta}}{1-\frac{1}{2}e^{\vartheta}}\right)^2 = \left(2e^{-\vartheta} - 1\right)^{-2}$$

$$M'(\vartheta) = -2\left(2e^{-\vartheta} - 1\right)^{-3} \cdot -2e^{-\vartheta} = 4\left(2e^{-\vartheta} - 1\right)^{-3} e^{-\vartheta}$$

dus $\mu = E(X) = M'(0) = 4(2-1)^{-3} = \boxed{4}$

$$M''(\vartheta) = -12\left(2e^{-\vartheta} - 1\right)^{-4} \cdot -2e^{-\vartheta} \cdot e^{-\vartheta} + 4\left(2e^{-\vartheta} - 1\right)^{-3} \cdot -e^{-\vartheta}$$

dus $E(X^2) = M''(0) = 24 - 4 = \boxed{20}$

zodat $\sigma^2 = 20 - 4^2 = \boxed{4}$

***Mogelijkheid 2**: X heeft overaftelbaar veel waarden.*

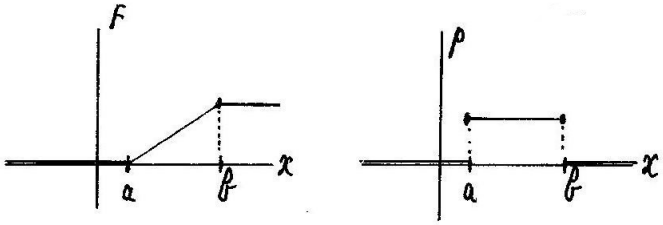
Hoe moet de verdelingsfunctie F(x) nu worden gedefiniëerd?

<u>Voorbeeld</u>
Ω overaftelbaar.
Zie par. 17.
Experiment: kies het punt P aselect op het interval [a,b].
Dan Ω = [a,b], dus Ω is op te vatten als de verzameling waarden op een gesloten interval.
Zij X de getrokken waarde.
In tegenstelling tot voorgaande voorbeelden is X hier niet <u>discreet</u> verdeeld (d.w.z. heeft hoogstens aftelbaar oneindig veel waarden), maar <u>continu</u> verdeeld en is sprake van een zogenaamde <u>kansdichtheid</u>.
Naam van deze verdeling: <u>uniforme kansverdeling</u>.
De definitie van verdelingsfunctie is analoog aan de definitie voor een dicreet verdeelde toevalsvariabele en luidt nu:

$F(x) = P(X \leq x) = 0$ als $x \leq a$
$= \frac{x-a}{b-a}$ als $a \leq x \leq b$
$= 1$ als $b \leq x$.

Hieruit volgt dat $F(x) = \int_{-\infty}^{x} p(t)\,dt$ met $p(t) = \frac{1}{b-a}$ als t tussen a en b ligt en elders 0.

De functie p heet <u>kansdichtheid</u>.
De analogie wordt ook zichtbaar als we bij F denken aan een oppervlaktebepaling.
De functie F is continu, de functie p niet, zoals hieronder is te zien.
De verwachtingswaarde en standaardeviatie zijn ook weer op de bekende wijzen te berekenen.



F continu ρ niet continu

$$\mu = E(X) = \int_{-\infty}^{\infty} x\,\rho(x)\,dx = \int_a^b x \cdot \frac{1}{b-a}\,dx = \frac{1}{b-a}\,\frac{1}{2}x^2\Big|_{x=a}^{x=b} = \frac{\frac{1}{2}(b^2-a^2)}{b-a} = \boxed{\frac{1}{2}(a+b)}$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2\,\rho(x)\,dx = \int_a^b (x-\mu)^2 \cdot \frac{1}{b-a}\,dx = \frac{\frac{1}{3}(x-\mu)^3}{b-a}\Big|_a^b = \frac{(b-\mu)^3-(a-\mu)^3}{3(b-a)} = \frac{\left\{\frac{1}{2}(b-a)\right\}^3 - \left\{\frac{1}{2}(a-b)\right\}^3}{3(b-a)} = \frac{\frac{1}{4}(b-a)^3}{3(b-a)}$$

$\mu = \frac{1}{2}(a+b)$

$$= \boxed{\frac{1}{12}(b-a)^2}$$

Momenten

$$M(\vartheta) = E\left(e^{\vartheta X}\right) = \int_{-\infty}^{\infty} e^{\vartheta x}\,\rho(x)\,dx = \int_a^b e^{\vartheta x} \cdot \frac{1}{b-a}\,dx = \frac{1}{b-a}\,\frac{1}{\vartheta}\,e^{\vartheta x}\Big|_a^b = \frac{1}{b-a}\,\frac{1}{\vartheta}\left(e^{\vartheta b} - e^{\vartheta a}\right)$$

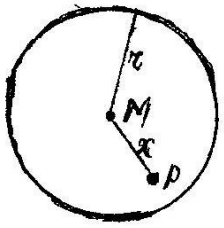
$$= \frac{1}{b-a}\,\frac{1}{\vartheta}\left(1+\vartheta b+\frac{1}{2}\vartheta^2 b^2+\frac{1}{6}\vartheta^3 b^3+\dots - 1-\vartheta a-\frac{1}{2}\vartheta^2 a^2-\frac{1}{6}\vartheta^3 a^3-\dots\right)$$

$$= \frac{1}{b-a}\left\{(b-a)+\frac{1}{2}\vartheta(b^2-a^2)+\frac{1}{6}\vartheta^2(b^3-a^3)+\dots\right\}$$

$$M'(0) = \boxed{\frac{1}{2}(b+a)} \quad \text{en} \quad M''(0) = \frac{1}{3}\,\frac{b^3-a^3}{b-a} = \boxed{\frac{1}{3}(b^2+ba+a^2)}$$

$$\sigma^2 = M''(0) - M'(0)^2 = \frac{1}{12}(b-a)^2$$

Voorbeeld
Ω overaftelbaar.
Zie par. 17.
Experiment: kies punt P aselect op het cirkeloppervlak van een cirkel met middelpunt M en straal r.
X; de afstand van punt P tot het middelpunt van de cirkel.



Verdelingsfuntie, verwachtingswaarde en standaarddeviatie:

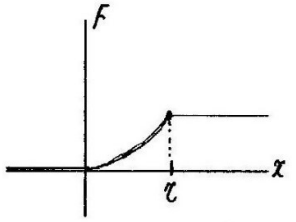
Dan $F(x) = P(X \leq x) = 0$ als $x \leq 0$

$= \frac{\pi x^2}{\pi r^2} = \frac{x^2}{r^2}$ als $0 \leq x \leq r$

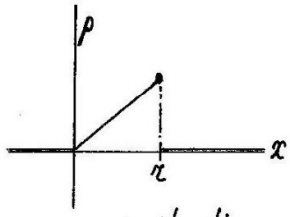
$= 1$ als $r \leq x$

en $F(x) = \int_{-\infty}^{x} \rho(t)\,dt$ met $\rho(x) = \frac{2x}{r^2}$ voor $x \in [0, r]$, elders 0

(kansdichtheid)



F continu



ρ niet continu

$$\mu = E(X) = \int_{-\infty}^{\infty} x\,\rho(x)\,dx = \int_{0}^{r} x \cdot \frac{2x}{r^2}\,dx = \frac{\frac{2}{3}x^3}{r^2}\Big|_0^r = \boxed{\frac{2}{3}r}$$

$$\sigma^2 = V(X) = \int_{0}^{r} x^2 \rho(x)\,dx - \mu^2 = \int_{0}^{r} \frac{2x^3}{r^2}\,dx - \mu^2 = \frac{\frac{1}{2}x^4}{r^2}\Big|_0^r - \mu^2 = \frac{1}{2}r^2 - \frac{4}{9}r^2 = \boxed{\frac{1}{18}r^2}$$

20. Normale verdelingen

Er zijn zeer veel toevalsexperimenten die (nagenoeg) overaftelbaar veel uitkomsten hebben. Hiervoor hebben we al enige voorbeelden laten zien. Maar er zijn er veel en veel meer.

Laat X: Ω → R een willekeurige toevalsvariabele zijn.
Onder de verdelingsfunctie van X wordt de volgende functie verstaan:

F(x) = P(X ≤ x) = P{ω ε Ω| X(ω) ≤ x}.

We beschouwen hier de situatie dat het aantal mogelijke waarden van X overaftelbaar is en dat F(x) = $\int_{-\infty}^{x} p(t)dt$ voor een zekere functie p.
In par. 19 hebben we bij mogelijkheid 2 al meerdere voorbeelden hiervan gezien.
Dan heet X continu verdeeld en de functie p heet kansdichtheid.
μ en σ zijn gedefiniëerd als μ = E(X) = $\int_{-\infty}^{\infty} xp(x)dx$ en
σ^2 = V(X) = $\int_{-\infty}^{\infty}(x-\mu)^2 p(x)dx$.

We bestuderen nu het bijzondere en zeer veel voorkomende geval:

p(t) = $\frac{1}{b\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-a}{b}\right)^2}$

Stelling
a = μ en b = σ

Bewijs:

$$E(X) = \int_{-\infty}^{\infty} x\,p(x)\,dx = \int_{-\infty}^{\infty} \frac{x}{b\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-a}{b}\right)^2} dx = \int_{-\infty}^{\infty} \frac{bz+a}{b\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \cdot b\,dz \qquad \left(z = \frac{x-a}{b}\right)$$

$$= \frac{b}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} \cdot z\,dz + \frac{a}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz = \frac{b}{\sqrt{2\pi}} \cdot -e^{-\frac{1}{2}z^2}\Big|_{-\infty}^{\infty} + \frac{a}{\sqrt{2\pi}} \cdot \sqrt{2\pi} = 0 + a$$

$$= a, \text{ dus } \underline{a = E(X) = \mu}$$

$$V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\,dx = \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{b\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{b}\right)^2} dx = \int_{-\infty}^{\infty} \frac{b^2z^2}{b\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \cdot b\,dz \qquad \left(z = \frac{x-\mu}{b}\right)$$

$$= \frac{b^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{1}{2}z^2} dz = \frac{b^2}{\sqrt{2\pi}} \sqrt{2\pi} = b^2, \text{ dus } \underline{b^2 = V(X) = \sigma^2(X)}$$

partiële integratie

Samengevat:

Als F(x) = P(X ≤ x) te schrijven is als **F(x) =** $\int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$ dan heet de toevalsvariabele X normaal verdeeld of Gauss-verdeeld.
(DeMoivre, 1733)
Zeer veel toevalsvariabelen zijn normaal verdeeld, zoals we verderop zullen zien.
We geven een aantal voorbeelden:

- Het geboortegewicht van pasgeboren baby's
- De lengte van vowassen mannen
- Het percentage CO2 in de lucht boven de stad Almere
- Het percentage ondeugdelijke exemplaren in een afgeleverde zending
- De jaarlijkse melkproduktie van een bepaalde melkkoe
- Het percentage onvoldoende toetsen bij een omvangrijk jaarlijks examen
- Het percentage personen in stad A die tbc hebben
- Het percentage 18-jarigen in stad B die geen alcohol drinken
- Het percentage linkshandige kinderen op 12-jarige leeftijd
- Het gewicht van geoogste sinaasappels op proefveld X
-

Enz.

Bij veel discrete kansverdelingen wordt ook vaak gedaan alsof ze continu zijn, vooral bij grote aantallen.

De kansdichtheid p is wel degelijk een kansverdeling, want:

$$\int_{-\infty}^{\infty} p(x)\,dx = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{2}{\sigma\sqrt{2\pi}}\int_{0}^{\infty} e^{-\frac{1}{2}z^2}\cdot\sigma\,dz = 1$$

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

Dat de toevalsvariabele X normaal verdeeld is betekent het volgende:

$$P(X<x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}s^2}\cdot\sigma\,ds = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}s^2} ds = \Phi(z)$$

$$s = \frac{t-\mu}{\sigma} \qquad z = \frac{x-\mu}{\sigma}$$

zie hiervoor

<u>Stelling</u>
Voor de functie Φ bestaat een eenvoudige en goede benadering:

$\Phi(z) \approx \frac{e^{1,7\,z}}{1+e^{1,7\,z}}$ (Molenaar, 1974)

Zij X een normaal verdeelde toevalsvariabele. Dan is de momentvoortbrengende functie ook te berekenen:

$$M(\vartheta) = e^{\mu\vartheta + \frac{1}{2}\sigma^2\vartheta^2}$$

Bewijs:

$$M(\vartheta) = E\left(e^{\vartheta X}\right) = \int_{-\infty}^{\infty} e^{\vartheta x} p(x)\,dx = \int_{-\infty}^{\infty} e^{\vartheta x} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left\{\frac{x}{\sigma} - \sigma\left(\vartheta + \frac{\mu}{\sigma^2}\right)\right\}^2 + \frac{1}{2}\sigma^2\left(\vartheta + \frac{\mu}{\sigma^2}\right)^2 - \frac{1}{2}\frac{\mu^2}{\sigma^2}} dx$$

$$= e^{\mu\vartheta + \frac{1}{2}\sigma^2\vartheta^2} \times \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}t^2} \cdot \sigma\,dt = e^{\mu\vartheta + \frac{1}{2}\sigma^2\vartheta^2} \cdot 1 \quad \text{qed} \qquad t = \frac{x}{\sigma} - \sigma\left(\vartheta + \frac{\mu}{\sigma^2}\right)$$

M.b.v. hiervan zijn ook weer μ en σ te berekenen.

$$M'(\vartheta) = e^{\mu\vartheta + \frac{1}{2}\sigma^2\vartheta^2}\left(\mu + \sigma^2\vartheta\right), \text{ dus } E(X) = M'(0) = e^0(\mu + 0) = \boxed{\mu}$$

$$M''(\vartheta) = e^{\mu\vartheta + \frac{1}{2}\sigma^2\vartheta^2}\left\{\left(\mu + \sigma^2\vartheta\right)^2 + \sigma^2\right\}, \text{ dus } E(X^2) = M''(0) = e^0\left(\mu^2 + \sigma^2\right) = \mu^2 + \sigma^2$$

$$\text{zodat } V(X) = \left(\mu^2 + \sigma^2\right) - \mu^2 = \boxed{\sigma^2}$$

<u>Stelling</u>
Als te toevalsvariabele X normaal verdeeld is, dan heeft de verdelingsfunctie F(x) een buigpunt voor x = μ en de kansdichtheid p(x) een buigpunt voor x = μ + σ en x = μ – σ.

Bewijs:

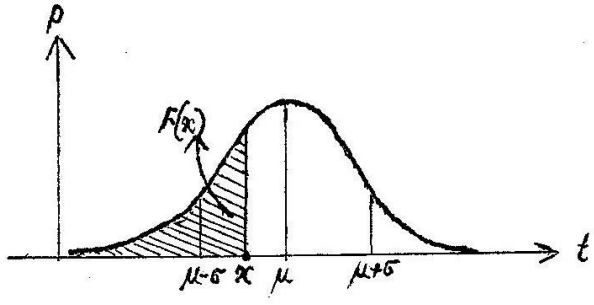
$$F'(x) = p(x)$$

$$F''(x) = p'(x) = p(x) \cdot -\frac{x-\mu}{\sigma} \cdot \frac{1}{\sigma} = p(x)\frac{\mu - x}{\sigma^2}$$

dus $F''(x) > 0$ voor $x < \mu$ en $F''(x) < 0$ voor $x > \mu$, zodat de grafiek van F in $x = \mu$ van hol naar bol gaat.

$$p''(x) = p'(x)\frac{\mu - x}{\sigma^2} + p(x) \cdot \frac{-1}{\sigma^2} = p(x)\frac{\mu - x}{\sigma^2}\frac{\mu - x}{\sigma^2} - p(x) \cdot \frac{1}{\sigma^2} = \frac{p(x)}{\sigma^4}\left\{(\mu - x)^2 - \sigma^2\right\} = 0 \text{ als } x - \mu = \pm\sigma$$

Karakteristiek voor de normale verdeling is de typische <u>klokvorm</u>:



<u>Stelling</u>
Als X normaal verdeeld is, dan ook Y = aX +b.

Bewijs:

Is vrij eenvoudig uit de definitie af te leiden.

Stelling
Als de toevalsvariabelen X en Y onafhankelijk en beide normaal verdeeld zijn, dan is ook X+Y normaal verdeeld.

Bewijs:
Wordt achterwege gelaten.
Vanwege de onafhankelijkheid geldt $E(XY) = E(X)E(Y)$ en $V(X+Y) = V(X) + V(Y)$.

Centrale limietstelling
De som van een groot aantal onafhankelijke toevalsvariabelen is bij benadering normaal verdeeld.

Bewijs wordt achterwege gelaten.
Vanwaar deze naam?
Omdat het om een groot aantal toevalsvariabelen gaat ($n \to \infty$) en bij een normale verdeling de waarden van de som dichter naar het midden (μ) kruipen.

Stelling
Bij de binomiale verdeling is de toevalsvariabele S_n normaal verdeeld, mits $npq > 9$.

Bewijs:
$S_n = X_1 + \ldots\ldots + X_n$ met X_i onafhankelijk. Pas nu de centrale limietstelling toe.

21. Steekproeven

Laat Ω bestaan uit overaftelbaar veel mogelijke uitkomsten en zij X een toevalsvariabele.
Wat als de belangrijke parameters μ = E(X) en σ^2 = V(X) niet bekend zijn?
Dan is een steekproef nodig om deze te kunnen schatten.
Met of zonder teruglegging? Bij een grote steekproef maakt dit weinig uit. Dus dan komt de steekproef neer op trekking met teruglegging.
Als n de steekproefgrootte is, dan kunnen we de volgende toevalsvariabelen definiëren:

X_i: de uitkomst (of waarde) van de i^e trekking, i = 1 t/m n

$\underline{X} = \frac{X_1+\cdots+X_n}{n}$

Voor grote n komt de steekproef neer op aselecte trekking met teruglegging en dan zijn de X_i's als onafhankelijk te zien en dus ook net zo verdeeld zijn als X. Dus $E(X_i) = E(X) = \mu$ en $V(X_i) = V(X) = \sigma^2$.

Hieruit volgt: $E(\underline{X}) = \frac{1}{n}\sum_i E(X_i) = \frac{n\mu}{n} = \mu$ vanwege de lineariteit van E.
$\underline{X}$ heet dan een zuivere schatting van μ. Als de steekproef groot genoeg is, dan is het gemiddelde van de steekproef bijna zeker een goede benadering van het populatiegemiddelde.

En ook: $V(\underline{X}) = \frac{1}{n^2}\sum_i V(X_i) = \frac{1}{n^2}\sum_i \sigma^2 = \frac{1}{n^2}.n\sigma^2 = \frac{\sigma^2}{n}$ vanwege de onafhankelijkheid van de X_i's.
Dus $V(\underline{X})$ is willekeurig klein te krijgen door n groot genoeg te nemen.
Dus bewezen:

$E(\underline{X}) = \mu$ en $V(\underline{X}) = \frac{\sigma^2}{n}$

Als X normaal verdeeld is, dan zijn alle X_i en ook $\underline{X}$ normaal verdeeld.
Een opmerkelijk en belangrijk gevolg van de centrale limietstelling is echter de volgende stelling:

Als X niet normaal verdeeld is, dan is voor grote n (n ≥ 30) $\underline{X}$ ook normaal verdeeld.

Een schatting van σ^2 is te geven door $s^2 = \frac{1}{n-1}\sum (X_i - \underline{X})^2$.

Stelling
$E(s^2) = \sigma^2$.

Bewijs:

$$\sum (X_i - \underline{X})^2 = \sum \{(X_i-\mu)-(\underline{X}-\mu)\}^2 = \sum (X_i-\mu)^2 - 2(\underline{X}-\mu)\sum (X_i-\mu) + n(\underline{X}-\mu)^2$$
$$= \sum (X_i-\mu)^2 - 2(\underline{X}-\mu)(n\underline{X}-n\mu) + n(\underline{X}-\mu)^2 = \sum (X_i-\mu)^2 - n(\underline{X}-\mu)^2$$
$$\text{Dus } E(s^2) = \frac{1}{n-1} E\sum (X_i-\underline{X})^2 = \frac{1}{n-1} E\left\{\sum (X_i-\mu)^2 - n(\underline{X}-\mu)^2\right\}$$
$$= \frac{1}{n-1}\sum \left\{E(X_i-\mu)^2 - nE(\underline{X}-\mu)^2\right\} = \frac{1}{n-1}\left\{\sum V(X_i) - nV(\underline{X})\right\} = \frac{1}{n-1}(n\sigma^2 - \sigma^2)$$
$$= \sigma^2$$

Dus ook s^2 is een zuivere schatting van σ^2.

Stelling

$V(s^2) = \frac{2\sigma^4}{n-1}$

Geen afleiding.
Ook $V(s^2)$ is dus willekeurig klein te krijgen door n groot genoeg te nemen.

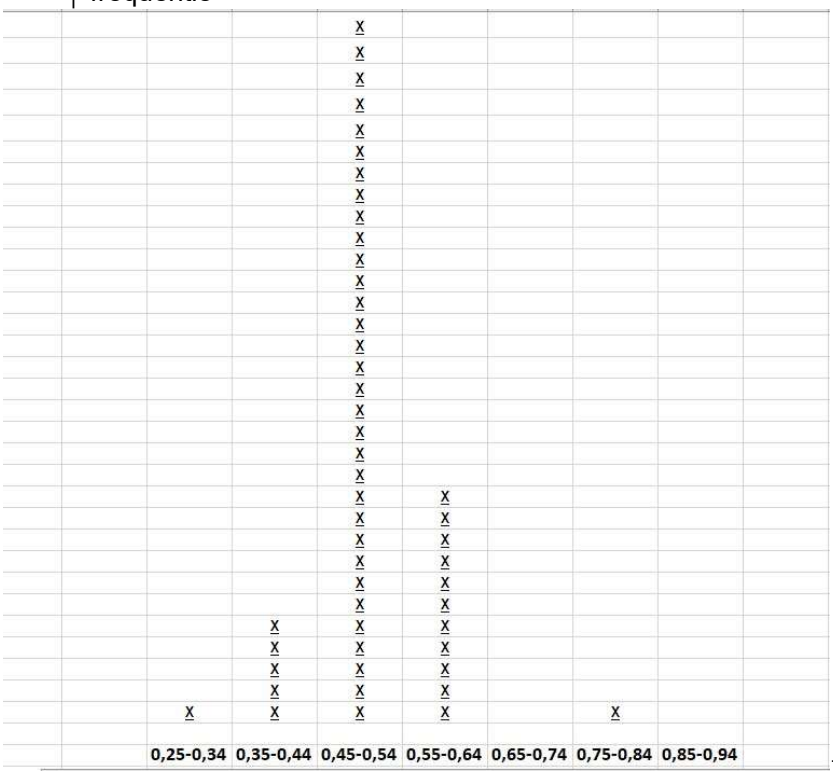
Ga nu zelf eens het volgende na:
Bij een steekproevenverdeling is de steekproefproportie $\hat{p}$ bij een voldoende grote steekproef bij benadering normaal verdeeld. Het gemiddelde van $\hat{p}$ is gelijk aan de populatieproportie p. Een goede schatting van de standaardafwijking van de populatieproportie p is:

$$\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}$$

Bij toename van de steekproefomvang neemt de geschatte spreiding dus af.

Een voorbeeld dat het steekproefgemiddelde normaal verdeeld is, voor n ≥ 30, ook als de populatie dat niet is:
Experiment: kies een punt P aselect op het interval [a,b]. Zie par. 19.
Dan Ω = [a,b], dus Ω is op te vatten als de verzameling waarden op een gesloten interval.
Zij X de getrokken waarde, dan is X continu verdeeld.
Naam van deze verdeling: uniforme kansverdeling.
Dit experiment is prachtig te simuleren met de Monte Carlo methode.
Het interval kunnen we opvatten als het interval [0,1]. Een willekeurig punt op dit interval komt neer op een decimaal getal tussen 0 en 1. Zulke getallen kunnen we aselect genereren in Excel m.b.v. het commando = ASELECT(). Door een kolom van n velden te vullen met n van deze decimale getallen simuleren we een steekproef ter grootte van n. Van deze n decimale getallen berekenen we nu het gemiddelde X m.b.v. Excel.
We voeren 50 keer deze steekproef uit. Voor n hebben we niet minstens 30 genomen, maar 27, omdat toen het spreadsheet nog mooi op het scherm paste. Zo hebben we dus 50 steekproefgemiddelden X verkregen.
Zie hier de frequentieverdeling van deze 50 steekproefgemiddelden:

↑ frequentie



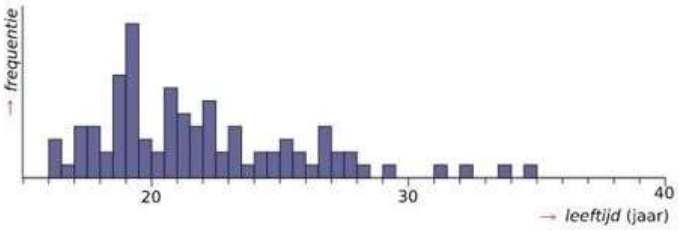
⟶ dec. getal

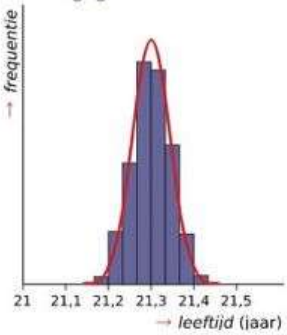
Toelichting: de tweede kolom geeft aan dat het steekproefgemiddelde 5 keer een waarde tussen 0,35 en 0,44 heeft.
De klokvorm verschijnt hier al duidelijk, na 50 steekproeven ter groote van 27.
De lezer kan zelf zelf de simulatie uitbreiden door grotere waarden voor 50 en 27 te nemen om dan te zien of de klokvorm nog mooier wordt.

Een ander voorbeeld, gevonden op internet, waarvan ik twijfels heb of het hier wel om echt uitgevoerde steekproeven gaat. De verwerking van de bedachte uitkomsten demonstreert echter wel duidelijk dat het steekproefgemiddelde normaal verdeeld is, ook als de populatie evident niet normaal verdeeld is.

Er is een heel groot concert met tienduizenden bezoekers. De organisatoren van het concert willen de gemiddelde leeftijd van de bezoekers weten.

Bij elk van de 50 ingangen zetten ze een enquêteur die aan elke 10^e bezoeker de leeftijd vraagt. Zo worden er 50 steekproeven genomen. Ga ervan uit dat deze steekproeven representatief zijn. Omdat niet iedereen wordt ondervraagd, kun je de gemiddelde leeftijd niet precies te weten komen. Je kunt deze alleen maar schatten. In het linker histogram zijn de gegevens van één van de 50 steekproeven weergegeven.





Het lijkt erop dat de leeftijden van de concertbezoekers niet normaal verdeeld zijn. Van alle 50 steekproeven die genomen zijn, is de gemiddelde leeftijd berekend, bekijk het rechter histogram. Het lijkt er op dat de steekproevenverdeling wel bij benadering normaal verdeeld is.

22. Axiomatische opbouw van de kansrekening

Het eerste boek over kansrekening is van Huygens: “De ratiociniis in ludo alea”, 1657, een verhandeling over de vragen bij kansspelen. Een aanzet was al eerder gegeven door Pascal en Fermat in hun befaamde briefwisselingen van begin 1654. Pas in de 17^{e} eeuw begon de kansrekening zich te ontwikkelen, hoewel reeds 3500 v. Chr. vragen rond bepaalde dobbelspellen oprezen.
De eerste die probeerde het begrip kans te definiëren was Laplace (1812). Zie par. 1. Aan deze definitie kleven wel enige bezwaren. Wat betekent onderling fysisch gelijkwaardig? Toch niet dat de kansen gelijk zijn? Maar ook, dat deze definitie onbruikbaar is als er oneindig veel mogelijke uitkomsten zijn.
De volgende definitie van kans werd begin van de 20^{e} eeuw door Von Mises (1920) gehanteerd. Zie par. 12.3 voor deze zgn. frequentiedefinitie, die gestoeld is op de ervaringswet van de grote aantallen. Hier kleven in wiskundig opzicht nog meer bezwaren aan.

Gedurende lange tijd werd kansrekening bedreven op grond van experimenten met een eindig aantal even waarschijnlijke uitkomsten. Op tamelijk gekunstelde wijze werden situaties die niet direct op deze wijze beschreven konden worden, zo gemodelleerd dat zij toch in dit raamwerk pasten. Meer en meer leidde dit tot onoverkomelijke moeilijkheden in de theorie. In 1933 publiceerde Kolmogorov het leerboek *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Daarmee doorbrak hij de impasse door een axiomatische aanpak van de kansrekening voor te stellen en daarmee een strenge onderbouwing van de kansrekening te geven

De kansfunctie is niet alltijd voor alle deelverzamelingen van de uitkomstenverzameling Ω gedefiniëerd, maar vaak alleen voor een bepaalde deelcollectie van deelverzamelingen. Die deelverzamelingen worden dan gebeurtenissen genoemd.
Op basis van de axioma’s kunnen dan alle bekende stellingen worden afgeleid. Deze axiomatische opbouw wordt hier verder niet bestudeerd. We laten alleen de axioma’s zien die de basis vormen van dit bouwwerk.

Als bij een bepaalde collectie deelverzamelingen $\mathbb{C}$ van een verzameling Ω een afbeelding P is gedefiniëerd die aan elke deelverzameling A in deze collectie een getal P(A) toekent met de volgende eisen:

1. $P(A) \geq 0$ voor elke A uit deze collectie,
2. $P(\Omega) = 1$,
3. $P(A \cup B) = P(A) + P(B)$ als $A \cap B = \emptyset$,

3a. $P(\bigcup_i A_i) = \sum_i P(A_i)$ voor onderling disjunctie A_i ($i=1,2,\ldots$) uit deze collectie,

dan heet het tripel Ω, $\mathbb{C}$, P een kansruimte en P een kansfunctie.
$\mathbb{C}$ is dus een bepaalde collectie deelverzamelingen van Ω. Deze deelverzamelingen worden de gebeurtenissen genoemd.

Dit is echter niet vodoende. De collectie $\mathbb{C}$ van deelverzamelingen moet ook nog aan het volgende voldoen:

a. Ω moet zelf tot de collectie behoren, dus $\Omega \in \mathbb{C}$.
b. Als $A \in \mathbb{C}$, dan ook $\neg A \in \mathbb{C}$. Dus $\mathbb{C}$ moet gesloten zijn m.b.t. complementvorming.
c. Als $A_i \in \mathbb{C}$ voor $i = 1,2,\ldots$ dan ook $\bigcup_i A_i \in \mathbb{C}$.

Een dergelijke collectie deelverzamelingen heet een σ-algebra.

Gevolgen
Ook $\emptyset \in \mathbb{C}$, want de lege verzameling is het complement van Ω.
Axioma 3a geldt ook voor een eindige vereniging van onderling disjuncte deelverzamelingen uit de collectie $\mathbb{C}$. Dit is met volledige inductie te bewijzen.

$\mathbb{C}$ is niet alleen gesloten m.b.t. willekeurige vereniging, maar ook m.b.t. willekeurige doorsnede. Dit laatste volgt uit $\cap_i A_i = \neg(U_i \neg A_i)$.

Stelling
$P(\emptyset) = 0$

Bewijs:
$P(\Omega) = P(\Omega \cup \emptyset) = P(\Omega) + P(\emptyset)$ volgens axioma 3.

Stelling
$P(\neg A) = 1 - P(A)$

Bewijs:
$1 = P(\Omega) = P(A \cup \neg A) = P(A) + P(\neg A)$ volgens axioma 3.
Stelling
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Bewijs:
Ga dit zelf na.

Enz.

Zo zijn alle bekende stellingen af te leiden uitgaande van de 4 genoemde axioma's.

We willen nu nog laten zien dat in de hiervoor bestudeerde stukken kansrekening inderdaad sprake is van kansruimtes volgens de hier gegeven definitie.

Toevalsexperimenten met een eindig aantal mogelijke uitkomsten
Als Ω uit n mogelijke uitkomsten bestaat, dan zijn er 2^n deelverzamelingen.
$\mathbb{C}$ is de collectie van alle deelverzamelingen van Ω.
Voor elke deelverzameling A is een kans P(A) gedefiniëerd en wel als volgt:
$P(A) = \sum_{\omega_i \epsilon A} p(\omega_i)$.
Er is dus zelfs een kans voor elke mogelijke uitkomst gedefiniëerd.
Het tripel Ω, $\mathbb{C}$, P voldoet aan de gestelde eisen voor een kansruimte, zoals snel te zien is.

Toevalsexperimenten met aftelbaar oneindig veel mogelijke uitkomsten
Par. 16 voorbeeld 1:
Werpen met een zuivere dobbelsteen tot een zes valt.
Mogelijke uitkomsten: 6, ≠6 6, ≠6 ≠6 6, ……
Kansen resp.: 1/6, 5/6x1/6, $(5/6)^2$x1/6, ……..
$\mathbb{C}$ bestaat uit alle mogelijke deelverzamelingen A van Ω.
$P(A) = \sum_{\omega \varepsilon A} p(\omega)$
Aan alle eisen voor een kansruimte is voldaan. We controleren alleen nog axioma 3a:
$P(U_i A_i) = \sum_{\omega \in U_i A_i} p(\omega) = \sum_{\omega \in A_1} p(\omega) + \sum_{\omega \in A_2} p(\omega) + \ldots\ldots = P(A_1) + P(A_2) + \ldots\ldots$
omdat in een convergente reeks met niet-negatieve termen de volgorde van de termen gewijzigd mag worden en de termen in eindig of oneindig veel groepjes tezamen genomen mogen worden zonder dat de som van de reeks verandert.

Als Ω uit aftelbaar oneindig veel mogelijke uitkomsten bestaat, dan geldt algemeen: als $p_1, p_2, \ldots\ldots$ niet-negatieve getallen zijn met $\sum_i p_i = 1$, dan is er precies één kansfunctie waarvoor geldt $P(\{\omega_i\}) = p_i$.

Toevalsexperimenten met overaftelbaar veel mogelijke uitkomsten

In dit geval wordt de problematiek ingewikkelder. In par. 17 hebben we enige voorbeelden bestudeerd en deze laten zien dat geschikte deelverzamelingen van Ω (geschikt betekent geschikt voor toekenning van een kans) al gauw met de lengte van een zeker interval of met de oppervlakte van een zeker gebied te maken hebben. Maar wat houdt de definitie van oppervlakte nu precies in? Dan komen we terecht in het gebied van meetbare verzamelingen en dat valt onder maattheorie. Een theorie die hier niet verder behandeld zal worden.
De voorbeelden in par. 17 hebben wel laten zien dat een duidelijke toekenning van lengte of oppervlakte kan leiden tot zinvolle toekenning van kansen en dat het dus niet nodig is om de axionmatische opbouw van de kansrekening volledig te kennen.
Tot slot: het zwaartepunt van de kansrekening is overduidelijk gelegen in de studie van toevalsexperimentern met eindig veel mogelijke uitkomsten. De meeste toepassingen in de statistiek hebben dan ook betrekking op deze categorie.

23. Het hanteren van het begrip kans buiten de wiskunde

In het voorgaande hebben we gezien dat het begrip kans een lange en vaak ook moeilijke ontwikkeling heeft doorgemaakt in de geschiedenis van de wiskunde. Het is begonnen met het bestuderen van experimenten met een eindig aantal mogelijke uitkomsten. Hier manifesteerden zich al snel dilemma's, die wiskundigen van die tijd voor een raadsel plaatsten. Vervolgens kwamen experimenten met aftelbaar oneindig veel mogelijke uitkomsten aan bod. Maar pas in de 20^{e} eeuw ontstond een dieper inzicht in de kansrekening, toen men geconfronteerd werd met de vraag wat men aan moest met experimenten die zelfs overaftelbaar veel mogelijke uitkomsten kenden. Toen bleek dat het begrip kans niet meer bij elke mogelijke uitkomst zinvol te definiëren was, maar dat men zich moest beperken tot bepaalde deelverzamelingen van de uitkomstenverzameling Ω. De Russische wiskundige Kolmogorov heeft hier een belangwekkende bijdrage geleverd.

Terugkijkend mogen we concluderen dat de beoefening van kansrekening alleen mogelijk is als sprake is van goed gedefiniëerde en welomschreven experimenten, waarbij ook duidelijk is wat onder de mogelijke uitkomsten wordt verstaan.

Buiten de context van wiskunde is dat echter wel even wat anders! Het woord kans wordt zeer veelvuldig gebruikt in de media, de politiek, diverse uiteenlopende discussies, enz. Maar ook in de rechtsgang. Iedereen denkt dat het begrip wel duidelijk is en alle discussianten nemen voetstoots aan dat de gebruikte vocabulaires ook vergelijkbaar of zelfs eensluidend zijn. Bijna niets is echter minder waar.

We geven hiervan 3 voorbeelden.

Voorbeeld 1

Art. NRC over gymnasia: Gymnasium-schaamte
(NRC van 19 februari 2022)

In de politiek en de media wordt met grote regelmaat gesteld dat er veel kansenongelijkheid bestaat, op diverse gebieden.
De toenemende aandacht voor gelijke kansen zet ook de discussie over gymnasia onder druk. In het genoemde artikel in de NRC komen veel rectoren van gymnasia aan het woord die hier duidelijke uitspraken over doen. De uitspraak die ikzelf het meest relevant en realistisch vind is:
"Kleinschalige scholen, met goede leraren en gemiddelde klassengroottes, waar leerlingen gezien en gekend worden – dat werkt."
(Christine Hylkema, rector van het Vossius Gymnasium in Amsterdam)

Dit standpunt pleit dus niet voor grootschaligheid in het onderwijs, ook niet voor brede en verlengde brugklassen en evenmin voor selectie op een later moment in de schoolloopbaan, maar juist voor categorale scholen.

Voorbeeld 2

Lucia de Berk
De zaak-Lucia de Berk betreft een rechtszaak tegen de Nederlandse verpleegkundige Lucia de Berk, die verdacht werd van meerdere moorden. De Berk, in de Nederlandse media

aangeduid als Lucia de B., werd in maart 2003 door de rechtbank in Den Haag veroordeeld tot een levenslange gevangenisstraf.
Na de heropening van de zaak in 2008 kwam de rechter tot de conclusie dat de veroordeling een justitiële dwaling was. De Berk werd vrijgesproken in 2010.

Hoewel er oorspronkelijk was uitgegaan van meerdere misdrijven bleek uit de uiteindelijke uitspraak van het gerechtshof in Arnhem dat er geen enkel misdrijf had plaatsgevonden. De justitiële dwaling betrof niet de veroordeelde, maar de feiten, die ten onrechte als moord dan wel poging tot moord gekwalificeerd waren.

Bij haar veroordeling had de rechtbank onder andere gebruikgemaakt van statistische berekeningen. Deze werden later echter onderuit gehaald.

Voorbeeld 3

Notaris verraad Anne Frank

In 2022 verscheen het boek Het **verraad** van **Anne Frank**, waarin een theorie werd gepresenteerd dat een lid van de Joodse Raad het adres gekend zou hebben en dit verraden had, teneinde zijn eigen gezin te redden.
De verschijning van dit boek heeft voor een enorme storm van reacties in de media gezorgd. De kritiek spitst zich o.a. ook toe op de gehanteerde statistische werkwijze.

De ‘85 procent zekerheid’ waarmee een Joodse notaris is aangewezen als Anne Frank-verrader, is te stellig gebracht, zegt de man die de berekening maakte. Hoe dan ook is er kritiek op de toepassing van zulke statistiek op complete strafzaken.
Alkemade blijkt ongelukkig met de manier waarop het percentage in de publiciteit is gekomen. Dat getal werd door het CCT en de media gebracht als een ‘absolute kans’, terwijl hij het had gerapporteerd als een ‘voorwaardelijke kans’. „Het verschil tussen die twee is geen onbelangrijke subtiliteit”, zegt Alkemade. „De voorwaardelijke uitspraak ‘*als* deze getuige de waarheid spreekt, *dan* is de kans 85 procent dat verdachte schuldig is”, heeft een fundamenteel andere betekenis dan de absolute uitspraak ‘De kans is 85 procent dat verdachte schuldig is’.”

Alkemade heeft de afgelopen tien jaar in opdracht van rechtbanken, het openbaar ministerie en advocaten van verdachten geregeld rapporten opgesteld over de bewijswaardering in strafzaken met behulp van zogeheten Bayesiaanse analyses.