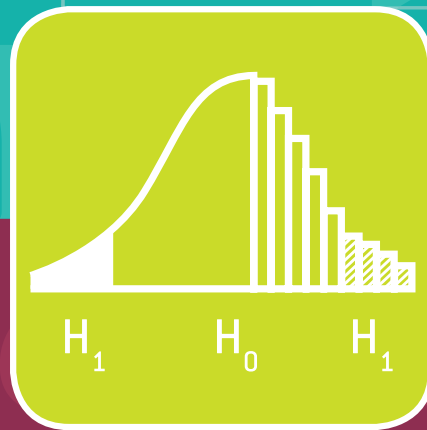


Jaap Klouwen



Statistische toetsen in economische toepassingen

uitgeverij
coutinho

STATISTISCHE TOETSEN IN ECONOMISCHE TOEPASSINGEN

Voor mijn ouders

Voor Reina, Bart, Koen en Anne

Statistische toetsen in economische toepassingen

Jaap Klouwen

Tweede, herziene druk

u i t g e v e r i j | C
c o u t i n h o

bussum 2016

www.coutinho.nl/statistischetoetsen

Je kunt aan de slag met het online studiemateriaal bij dit boek. Dit materiaal bestaat uit enkele Excelbestanden, in de vorm van tabellen, data of toepasselijke berekeningswijzen. Docenten kunnen de uitwerkingen van de opgaven aanvragen.

© 2009/2016 Uitgeverij Coutinho bv

Alle rechten voorbehouden.

Behoudens de in of krachtens de Auteurswet van 1912 gestelde uitzonderingen mag niets uit deze uitgave worden verveelvoudigd, opgeslagen in een geautomatiseerd gegevensbestand, of openbaar gemaakt, in enige vorm of op enige wijze, hetzij elektronisch, mechanisch, door fotokopieën, opnamen, of op enige andere manier, zonder voorafgaande schriftelijke toestemming van de uitgever.

Voor zover het maken van reprografische verveelvoudigingen uit deze uitgave is toegestaan op grond van artikel 16 h Auteurswet 1912 dient men de daarvoor wettelijk verschuldigde vergoedingen te voldoen aan Stichting Reprorecht (Postbus 3051, 2130 KB Hoofddorp, www.reprorecht.nl). Voor het overnemen van (een) gedeelte(n) uit deze uitgave in bloemlezingen, readers en andere compilatiewerken (artikel 16 Auteurswet 1912) kan men zich wenden tot Stichting PRO (Stichting Publicatie- en Reproductierechten Organisatie, Postbus 3060, 2130 KB Hoofddorp, www.stichting-pro.nl).

Eerste druk 2009

Tweede, herziene druk 2016

Uitgeverij Coutinho

Postbus 333

1400 AH Bussum

info@coutinho.nl

www.coutinho.nl

Zetwerk: BoekenWijs, Almere

Omslag: studio Pietje Precies bv, Hilversum

Noot van de uitgever

Wij hebben alle moeite gedaan om rechthebbenden van copyright te achterhalen. Personen of instanties die aanspraak maken op bepaalde rechten, wordt vriendelijk verzocht contact op te nemen met de uitgever.

ISBN 978 90 469 0530 2

NUR 916

Voorwoord bij de tweede druk

In deze herziene, tweede druk is, mede op aanraden van enkele gebruikers, een hoofdstuk over enkelvoudige en meervoudige regressieanalyse toegevoegd. Enkelvoudige lineaire regressie was ook aan bod gekomen in het *Basisboek kwantitatieve methoden – Statistiek met Exceltoepassingen* (Coutinho, 2013), maar is als onderwerp met de uitbreiding naar meer variabelen een natuurlijke aanvulling op het onderzoek naar een mogelijke oorzaak-gevolgrelatie tussen variabelen.

Verder is er een aanvulling op het deel over steekproefgrootte opgenomen in hoofdstuk 2. De vraag naar de steekproefgrootte bij een ‘eindige populatie’ komt namelijk veel voor in de praktijk (‘Hoe groot moet in een enquête mijn steekproef onder de 160 werknemers van dit bedrijf zijn als de resultaten als betrouwbaar kunnen worden beschouwd?’).

‘Onderzoekend vermogen’ is een bekende term in opleidingsprofielen geworden. Longitudinaal onderzoek (in de tijd gemeten resultaten verklaren) en transversaal onderzoek (resultaten verkregen op één tijdstip duiden) spelen daar ook een rol in. Een voorbeeld van longitudinaal onderzoek is de al genoemde (lineaire) regressieanalyse. Transversaal onderzoek treedt bijvoorbeeld op in een enquête waarbij men een verband tussen man/vrouw en het hebben van schulden onderzoekt, of, meer algemeen, in studies met kruistabellen.

De opgaven zijn deels aangepast en aangevuld met nieuw, actueel materiaal. Hierbij is de oude volgorde aangehouden; de nieuwe opgaven staan dus achteraan. Op een enkele plaats is een oude opgave aangepast.

Verder is gekozen voor de Nederlandse naam van de speciale statistische functies in Excel. Sinds Excel 2010 zijn (vooral in het Nederlands) de namen van die functies gewijzigd. Als hulp zijn daarom in appendix 2 van alle in dit boek gebruikte Excelfuncties de nieuwe en oude Nederlandse naam, alsmede de Engelse naam in een tabel opgenomen.

Redacteur Louise Prompers leverde uitstekend werk door veel onvolkomenheden in de kopij te signaleren. Ook bedank ik wiskunderedacteur Kim Fierens voor een aantal specifieke opmerkingen.

Als laatste bedank ik mijn (oud-)collega's René Debets, Dirk Gorter en Donald van As voor het aanleveren van nieuwe opgaven en/of het geven van commentaar op de eerste editie.

Amersfoort, april 2016
Jaap Klouwen

Inhoudsopgave

Inleiding	11
1 Kansverdelingen	13
1.1 Inleiding	13
1.2 Binomiale kansverdeling	13
1.3 Hypergeometrische kansverdeling	19
1.4 Normale kansverdeling	22
1.5 Poissonverdeling	28
1.5.1 Poissonformule	28
1.5.2 Enkele eigenschappen van de Poissonverdeling	33
1.6 De negatief-exponentiële kansverdeling	34
Opgaven	38
2 Betrouwbaarheidsintervallen	45
2.1 Inleiding	45
2.2 Normaal verdeeld betrouwbaarheidsinterval	45
2.3 Binomiaal verdeeld betrouwbaarheidsinterval	50
2.4 Betrouwbaarheidsinterval met de t -verdeling	51
2.5 Steekproefgrootte	55
2.6 Steekproefgrootte bij eindige populaties	57
2.6.1 Normaal verdeeld betrouwbaarheidsinterval	58
2.6.2 Binomiaal verdeeld betrouwbaarheidsinterval	59
2.7 Betrouwbaarheid bij gegeven N en n	62
Opgaven	64
3 Statistische toetsen	75
3.1 Inleiding	75
3.2 Binomiale toets op een percentage	75
3.3 Normale toets op een gemiddelde	81
3.4 t -toets op een gemiddelde	84
3.5 Poissontoets op een gemiddelde	86
3.6 χ^2 -toetsen	88
3.6.1 χ^2 -toets op een frequentietabel	88
3.6.2 χ^2 -toets op een kruistabel	94
3.7 Onderscheidingsvermogen en fouten van de eerste en tweede soort	99
Opgaven	105

4	Verschiltoetsen	119
4.1	Inleiding	119
4.2	Verschiltoetsen bij onafhankelijke waarnemingen	119
4.2.1	Standaardafwijkingen bekend	119
4.2.2	Standaardafwijkingen onbekend	122
4.3	Verschiltoetsen bij gepaarde waarnemingen	125
4.3.1	Tekentoets bij gepaarde waarnemingen	125
4.3.2	t -toets bij gepaarde waarnemingen	127
4.4	Verschiltoetsen bij proporties	129
	Opgaven	131
5	Controlekaarten	137
5.1	Inleiding	137
5.2	$\bar{X}$ -kaart en R -kaart bij bekende μ en σ	137
5.2.1	$\bar{X}$ -kaart	137
5.2.2	R -kaart	140
5.3	$\bar{X}$ -kaart en R -kaart bij onbekende μ en σ	141
5.4	p -kaart	145
5.5	c -kaart	147
	Opgaven	149
6	Mann-Whitneytoets	153
6.1	Inleiding	153
6.2	Mann-Whitneytoets voor kleine steekproeven	154
6.3	Mann-Whitneytoets voor grote steekproeven	157
	Opgaven	160
7	De waarde-eenhedensteekproef	163
7.1	Inleiding	163
7.2	Soorten steekproeven	163
7.3	Positieve en negatieve controle	165
7.4	De waarde-eenhedensteekproef	166
7.5	Toets van de waarde-eenhedensteekproef	167
7.6	Controle op ernstige fouten	169
	Opgaven	171
8	Enkelvoudige en meervoudige lineaire regressie	175
8.1	Inleiding	175
8.2	Regressielijn en correlatiecoëfficiënt	175
8.3	Regressie in Excel	180
8.4	Meervoudige regressie	183

8.5	Andere regressiemodellen	190
	Opgaven	196
Appendices		205
A1	Enkele statistische functies bij de grafische rekenmachine	206
A2	Lijst van gebruikte Excelfuncties, Nederlands ↔ Engels	210
A3	Standaardafwijking van de hypergeometrische en binomiale kansverdeling	212
A4	Benaderingen	213
A5	De R -kaart bij onbekende μ en σ	216
A6	Bewijs van enkele eigenschappen van de kansvariabele $\underline{R}_1$	218
Tabellen		219
Tabel 1	Enkele cumulatieve binomiale kansverdelingen	219
Tabel 2	Standaardnormale kansverdeling, z -tabel (rechteroverschrijdingskansen)	222
Tabel 2a	Waarden van z bij gegeven rechteroverschrijdingskans α	223
Tabel 3	Waarden van t bij gegeven rechteroverschrijdingskans α	224
Tabel 4	Enkele cumulatieve Poissonkansverdelingen	225
Tabel 5	Waarden van χ^2 bij gegeven rechteroverschrijdings- kans α	226
Tabel 6	Waarden van enkele factoren voor $\bar{X}$ - en R -controlekaarten	227
Tabel 7	Kritieke waarden Mann-Whitneytoets ($\alpha = 0,05$)	227
Literatuur		228
Register		229

Inleiding

Dit boek behandelt de basisbegrippen uit de statistische toetsingstheorie. Uit langdurige leservaring is mij gebleken dat studenten veel moeite hebben met het concept van een (klassieke) toets in de statistiek. Heel vaak wordt al snel – mijns inziens onterecht – een veelomvattend statistisch pakket als SPSS gebruikt. Naar mijn overtuiging dient eerst een theoretische, maar wel door praktische voorbeelden gestuurde benadering te worden gekozen. De praktische voorbeelden in dit boek komen uit de economie: bedrijfseconomie, bedrijfskunde, accountancy, logistiek en commerciële economie.

Dit boek beoogt een inleiding te zijn op de algemene toetsingstheorie; het pretendeert dus niet volledig te zijn. Ook gaat het boek niet over het trekken van steekproeven, net zomin als het gaat over methoden van enquêteren en het verzamelen van data.

Waar dit boek wel over gaat is het kunnen beoordelen van statistische conclusies zoals die voorkomen in krantenartikelen, op het internet en in vaktijdschriften. Bijvoorbeeld: mag een percentage uit een aselechte steekproef ‘doorgetrokken’ worden naar de gehele bevolking? Een ander voorbeeld: als een bedrijf beweert dat de kosten tussen de € 100.000 en € 120.000 per maand liggen, hoe betrouwbaar is deze mededeling dan? En, hoe vaak staat er niet in de krant dat een bepaald onderzoeksresultaat ‘significant’ is? Wat is precies ‘significant’?

Kort gezegd bestaat de basis van dit boek uit betrouwbaarheidsintervallen (ook wel ‘schatten van parameters’ geheten) en statistische toetsen. Deze onderwerpen komen aan bod in de eerste drie hoofdstukken. De vier daaropvolgende hoofdstukken zijn specifieke toetsen en/of toepassingen van de in de eerste alinea genoemde gebieden. Het laatste hoofdstuk gaat over het mogelijke verband tussen twee variabelen. Enkele technische bewijzen zijn in de appendices geplaatst.

Statistische toetsen in economische toepassingen kan door een hbo-student zelfstandig worden gebruikt. Het materiaal waarop dit boek is gebaseerd, is jarenlang gebruikt in hbo-opleidingen bedrijfseconomie, accountancy en logistiek. Ook in commercieel-economische opleidingen kan men met dit boek aan de slag, omdat de belangrijkste soorten toetsen worden besproken.

Uiteraard is het gebruik van een pc onontbeerlijk: er wordt getoond hoe Excel behulpzaam kan zijn bij allerlei statistische berekeningen. Ook wordt aandacht besteed aan het gebruik van de grafische rekenmachine. Op de website van het boek zijn enkele Excelbestanden te vinden, in de vorm van tabellen, data of toepasselijke berekeningswijzen.

De onderwerpen van dit boek sluiten aan bij die van het *Basisboek kwantitatieve methoden – Statistiek met Exceltoepassingen* (Coutinho, 2013). In dat boek worden onder andere de basisbegrippen van de kansrekening en de kansverdelingen binomiale en normale verdeling behandeld. Omwille van de bruikbaarheid als onafhankelijk boek wordt in het eerste hoofdstuk kort ingegaan op beide genoemde kansverdelingen. De aanpak van de materie sluit ook aan bij het *Basisboek statistiek*: een niet al te technische benadering, met voorbeelden uit het (dagelijkse) economische leven. Verder is geprobeerd niet te lang van stof te zijn, in een streven naar een compact boek. Mede daarom zijn de antwoorden niet in dit boek opgenomen, maar samen met uitwerkingen digitaal beschikbaar voor docenten, op aanvraag.

www.coutinho.nl/statistischetoetsen

Je kunt aan de slag met het online studiemateriaal bij dit boek. Dit materiaal bestaat uit enkele Excelbestanden, in de vorm van tabellen, data of toepasselijke berekeningswijzen.

Docenten kunnen de uitwerkingen van de opgaven aanvragen.

1 Kansverdelingen

1.1 Inleiding

In dit hoofdstuk worden drie kansverdelingen behandeld, naast de al bekend veronderstelde binomiale en normale kansverdeling (zie *Basisboek kwantitatieve methoden – Statistiek met Exceltoepassingen*, hoofdstukken 5 en 6). Beide kansverdelingen zijn voor de volledigheid ook kort opgenomen in dit boek.

De drie andere kansverdelingen zijn de kansverdeling die ontstaat bij steekproeftrekking ‘zonder teruglegging’ uit een populatie met twee items, ook de hypergeometrische kansverdeling genoemd (zoals de binomiale kansverdeling de kansverdeling ‘met teruglegging’ is); de Poissonverdeling, belangrijk bij wachttijdproblemen en accountantscontrole; en de negatief-exponentiële kansverdeling (de kansverdeling van de tijd tussen twee gebeurtenissen die Poisson verdeeld zijn).

1.2 Binomiale kansverdeling

In deze paragraaf wordt de binomiale kansverdeling kort besproken. Voor een volledige behandeling wordt verwezen naar het *Basisboek statistiek*. De binomiale verdeling is een bijzonder geval van een discrete kansverdeling, die haar economische toepassingen vooral binnen het marktonderzoek, de accountancy en de logistiek heeft. Denk daarbij aan percentagevraagstukken als verdeling van marktaandelen en kwaliteitscontrole. Belangrijk bij een binomiale kansverdeling is de aanname dat de kans op een bepaalde gebeurtenis constant is.

VOORBEELD 1.1

Stel dat de kans dat de AEX-index op een zekere beursdag stijgt, gelijk is aan $p = 0,6$. Stel vervolgens dat deze kans p gedurende een aantal beursdagen constant is en dat de beurs van dag tot dag onafhankelijk is. De kans dat de AEX-index op een zekere beursdag daalt of gelijk blijft is dus gelijk aan $1 - p = 1 - 0,6 = 0,4$, via de ontkenningregel (ook wel complementregel genoemd). Zie voor deze en andere regels uit de kansrekening hoofdstuk 3 uit het *Basisboek statistiek*. Stel verder dat we de Amsterdamse AEX 10 beursdagen volgen.

Hoe groot is de kans dat van die 10 dagen er 7 zijn geweest met een stijging? En hoe groot zijn de kansen op andere uitkomsten van het aantal stijgingen?

Oplossing:

De kans op eerst 7 dagen stijgen (S) en daarna 3 dagen dalen of gelijk blijven (D) is:
 $P(\text{ssssssDDDD}) = 0,6^7 \cdot 0,4^3 = 0,0018$

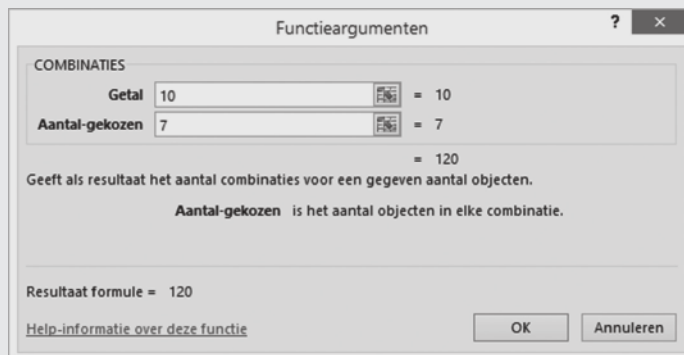
Maar sssssssDDDD is maar een van de vele manieren om 7 keer te stijgen en 3 keer te dalen. Een andere mogelijkheid (volgorde) is sssssDDDDs. Op hoeveel manieren kan men 7 keer S en 3 keer D opschrijven? Dit is de combinatie van 7 uit 10, geschreven als ${}_{10}C_7 = 120$. Met behulp van rekenapparatuur:

Excel: COMBINATIES(10;7)

NB De standaardfuncties in Excel beginnen steeds met het gelijkheidssymbool =. De cel in het voorbeeld bevat dus in werkelijkheid: =COMBINATIES(10;7). We zullen dit teken in dit boek verder weglaten.

We beschrijven hierna het stappenplan voor deze Excelfunctie.

- 1 Typ in een willekeurige cel in Excel: =COMBINATIES((Engels: COMBIN. In appendix 2 staat een lijst met alle in dit boek gebruikte Excel-functies, zowel in de vertaling Nederlands-Engels als andersom.)
- 2 Klik op de functiewizard: f_x
- 3 Er verschijnt een pop-upmenu. Vul hier bij Getal 10 in en bij Aantal-gekozen 7. Het antwoord 120 is dan al te zien:



Gewone rekenmachine: ${}_{10} \setminus nCr \setminus 7$

Texas grafische rekenmachine: ${}_{10} \setminus \text{MATH} \setminus \text{PRB} \setminus 3: nCr \setminus 7 \setminus \text{ENTER}$

Casio grafische rekenmachine: $\text{RUN} \setminus {}_{10} \setminus \text{OPTN} \setminus \blacktriangleright (F6) \setminus \text{PROB} (F3) \setminus nCr \setminus 7 \setminus \text{EXE}$

Als we de kansvariabele k definiëren als:

$\underline{k}$ = aantal dagen met een stijging

dan is de kans op 7 stijgingen van de 10:

$$P(\underline{k} = 7) = {}_{10}C_7 \cdot 0,6^7 \cdot 0,4^3 = 0,2150$$

Voorgaande berekeningswijze is een voorbeeld van de *binomiale formule*. Het systeem hierin is in principe eenvoudig. Wat is bijvoorbeeld de kans op 8 stijgingen van de 10?

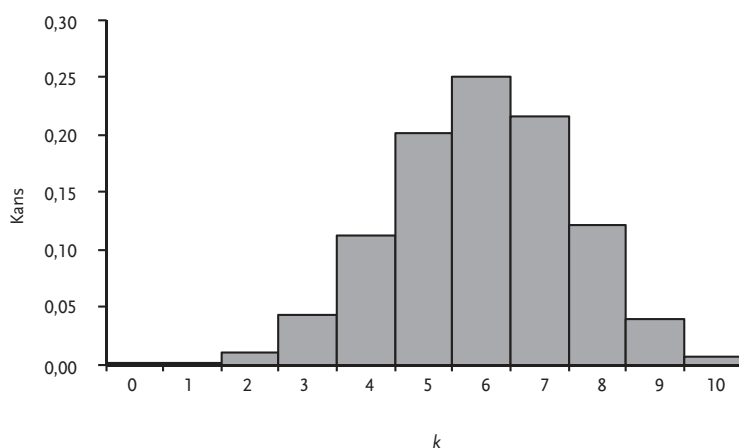
$$P(\underline{k} = 8) = {}_{10}C_8 \cdot 0,6^8 \cdot 0,4^2 = 0,1209 \quad (\text{NB } 8 + 2 = 10, \text{ in de vorige formule } 7 + 3 = 10, \text{ et cetera.})$$

Alleen de schuingedrukte cijfers zijn veranderd ten opzichte van de vorige formule. Zo is:

$$P(\underline{k} = 5) = {}_{10}C_5 \cdot 0,6^5 \cdot 0,4^5 = 0,2007$$

Rekenen we alle kansen uit (met mogelijke waarden van k van 0 tot en met 10) en maken we er een histogram van, dan krijgen we figuur 1.1. Dit is de kansverdeling van de binomiale verdeling met de twee *parameters* $n = 10$ en $p = 0,6$. Ook wel kort genoteerd als:

$$\underline{k} \sim \text{Bin}(10; 0,6)$$



Figuur 1.1 Histogram van de binomiale kansverdeling met $n = 10$ en $p = 0,6$

De *binomiaalformule* is:

$$P(\underline{k} = k) = {}_n C_k \cdot p^k \cdot (1 - p)^{n-k}$$

Binomiaal betekent dat er altijd *twee* beschrijvingen van het probleem mogelijk zijn. In het AEX-voorbeeld zijn dat het ‘aantal stijgende beursdagen’ en ‘aantal dalende of gelijkblijvende beursdagen’. De kans op 7 stijgende dagen is precies gelijk aan de kans op 3 dalende of gelijkblijvende dagen. Dit laatste is ook de reden waarom er in binomiale tabellen (zie tabel 1 achter in dit boek) kansen tot en met 0,5 worden opgenomen. Voor kansen groter dan 0,5 kan men immers overstappen op de complementaire gebeurtenis met eveneens complementaire kans.

Vervolgens bespreken we het gemiddelde en de standaardafwijking van de binomiale kansverdeling. Waarom heeft de kans bij $k = 6$ in het histogram van figuur 1.1 de grootste waarde? Het antwoord is: als we 10 dagen kijken, en elke dag is de kans op ‘succes’ 0,6, dan zal naar verwachting de uitkomst $10 \cdot 0,6 = 6$ het meest voorkomen. (Theoretisch is de modus (de waarde met de grootste kans) van een binomiale verdeling gelijk aan het grootste gehele getal dat kleiner is dan $(n + 1) \cdot p$.) Dit principe geldt algemeen. Het gemiddelde of de verwachtingswaarde (in het Engels *expectation*; notatie: μ of $E(\underline{k})$) van een binomiale verdeling $\underline{k}$ is:

$$\mu = n \cdot p$$

en de standaardafwijking is gelijk aan:

$$\sigma = \sqrt{n \cdot p \cdot (1 - p)}$$

Voor het AEX-voorbeeld is de standaardafwijking dus:

$$\sigma = \sqrt{10 \cdot 0,6 \cdot 0,4} = \sqrt{2,4} = 1,55$$

Het is belangrijk bij een binomiale kansverdeling te beseffen dat het een tel-kansverdeling is, met als mogelijke uitkomsten alleen 0, 1, 2, ... tot en met n (de steekproefgrootte). In de volgende hoofdstukken komen voornamelijk *cumulatieve* binomiale kansen voor. In het volgende voorbeeld wordt hierop ingegaan.

VOORBEELD 1.2

Een vierkeuzetoets bestaat uit 20 vragen, waarvan de antwoorden ‘op de gok’ zullen worden gegeven. We willen de volgende vier vragen beantwoorden:

- 1 Hoe groot is de kans op precies 10 goed gegokte antwoorden?
- 2 Hoe groot is de kans op ten hoogste 10 goed gegokte antwoorden?
- 3 Hoe groot is de kans op minstens 10 goed gegokte antwoorden?
- 4 Hoe groot is de kans op minstens 6 en hoogstens 10 goed gegokte antwoorden?

Oplossing:

- 1 We beantwoorden deze vraag met de drie al eerder gebruikte rekenmethoden:

Excel: $\text{BINOM.VERD}(10;20;0,25;0) = 0,009922$ (Oudere versies van Excel: BINOMIALE.VERD ; Engels: BINOMDIST .)

Functieargumenten

BINOM.VERD

Aantal-gunstig	10	= 10
Experimenten	20	= 20
Kans-gunstig	0,25	= 0,25
Cumulatief	0	= ONWAAR

= 0,009922275

Geeft als resultaat de binomiale verdeling.

Cumulatief is een logische waarde. Gebruik WAAR voor de cumulatieve verdelingsfunctie en ONWAAR voor de kansdichtheidsfunctie.

Resultaat formule = 0,009922275

[Help-informatie over deze functie](#)

OK Annuleren

Texas: $2^{\text{nd}} \setminus \text{DISTR} \setminus \text{binompdf}(20, 0.25, 10) = 0,009922$

Casio: $\text{STAT} \setminus \text{DIST} \setminus \text{BINM} \setminus \text{Bpd} \setminus \text{Data: Var} \setminus x: 10 \setminus \text{Numtrial: } 20 \setminus p: 0.25 \setminus \text{EXE}$ geeft 0,009922

De 0 (of ONWAAR; Engels: FALSE) in de Excelfunctie symboliseert een kans op precies 10. (Bij een 1 of TRUE zou een cumulatieve kans worden berekend; zie vraag 2.) Let erop dat bij de drie gebruikte soorten rekenapparatuur de volgorde verschillend kan zijn; Excel: $\text{BINOM.VERD}(k; n; p; 0 \text{ of } 1)$ versus Texas: $\text{binompdf}(n, p, k)$. In appendix 1 zijn uitgebreidere beschrijvingen voor de statistische functies van de twee meest gangbare typen grafische rekenmachines opgenomen.

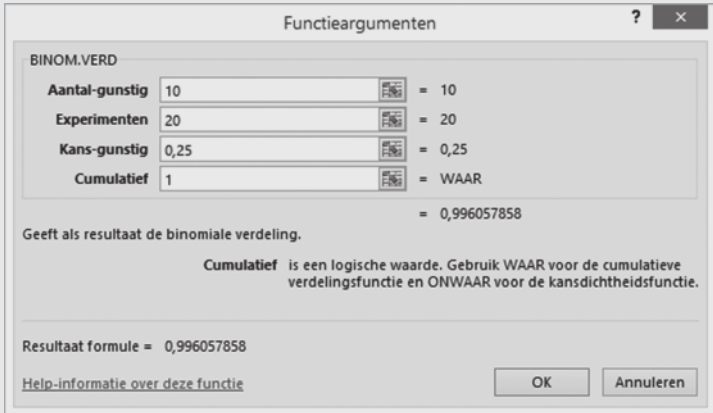
- 2 Als de kansvariabele $\underline{k}$ het aantal goed gegokte antwoorden voorstelt, zouden we voor de eerste vraag de kansen $P(\underline{k} = 0)$ tot en met $P(\underline{k} = 10)$ moeten optellen. De 'oude' manier hiervoor is het opzoeken van de kans $P(\underline{k} \leq 10)$ in een cumulatieve binomiale tabel. Achter in het boek, in tabel 1, zijn voor enkele waarden van n en p , cumulatieve tabellen opgenomen. Tabel 1.1 bevat een deel van de cumulatieve binomiale tabel voor $n = 20$ en verschillende waarden van p . Opzoeken bij $p = 0,25$ en $k = 10$ levert de gevraagde kans: $P(\underline{k} \leq 10) = 0,9961$.

Tabel 1.1 Deel van de cumulatieve binomiale tabel met $n = 20$ voor verschillende waarden van p

$n = 20$	p									
k	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
0	0,3585	0,1216	0,0388	0,0115	0,0032	0,0008	0,0002	0,0000	0,0000	0,0000
1	0,7358	0,3917	0,1756	0,0692	0,0243	0,0076	0,0021	0,0005	0,0001	0,0000
2	0,9245	0,6769	0,4049	0,2061	0,0913	0,0355	0,0121	0,0036	0,0009	0,0002
3	0,9841	0,8670	0,6477	0,4114	0,2252	0,1071	0,0444	0,0160	0,0049	0,0013
4	0,9974	0,9568	0,8298	0,6296	0,4148	0,2375	0,1182	0,0510	0,0189	0,0059
5	0,9997	0,9887	0,9327	0,8042	0,6172	0,4164	0,2454	0,1256	0,0553	0,0207
6	1,0000	0,9976	0,9781	0,9133	0,7858	0,6080	0,4166	0,2500	0,1299	0,0577
7	1,0000	0,9996	0,9941	0,9679	0,8982	0,7723	0,6010	0,4159	0,2520	0,1316
8	1,0000	0,9999	0,9987	0,9900	0,9591	0,8867	0,7624	0,5956	0,4143	0,2517
9	1,0000	1,0000	0,9998	0,9974	0,9861	0,9520	0,8782	0,7553	0,5914	0,4119
10	1,0000	1,0000	1,0000	0,9994	0,9961	0,9829	0,9468	0,8725	0,7507	0,5881
11	1,0000	1,0000	1,0000	0,9999	0,9991	0,9949	0,9804	0,9435	0,8692	0,7483

Tegenwoordig gebruiken we, zoals we al eerder zagen, Excel of de grafische rekenmachine voor het berekenen van dit soort kansen. Voor $P(k \leq 10)$ is dit:

Excel: BINOM.VERD(10;20;0,25;1)



Texas: 2nd \ DISTR \ binomcdf(20,0.25,10)

Casio: STAT \ DIST \ BINM \ Bcd \ Data: Var \ x: 10 \ Numtrial: 20 \ p: 0.25 \ EXE

(De 1 in de Excelfunctie symboliseert een – linkszijdige – cumulatieve kans.)

- 3 De kans op minstens 10 goed gegokte antwoorden is 1 minus de kans op hoogstens 9 goed gegokte antwoorden: $P(\underline{k} \geq 10) = 1 - P(\underline{k} \leq 9)$, via de ontkenningregel. Deze werkwijze is nodig, omdat zowel grafische rekenmachines als Excel alleen linkeroverschrijdingskansen kunnen berekenen bij de binomiale verdeling. Er volgt (op een soortgelijke manier als bij vraag 2): $1 - 0,9861 = 0,0139$.
- 4 Dit is de kans: $P(6 \leq \underline{k} \leq 10)$, als verschil van twee linkeroverschrijdingskansen te schrijven als: $P(\underline{k} \leq 10) - P(\underline{k} \leq 5)$. Voor beide kansen kan de methode uit antwoord 2 worden toegepast. Dit levert: $0,9961 - 0,6172 = 0,3789$.
Bij de vragen 3 en 4 kan eventueel ook de tabel weer worden gebruikt.

1.3 Hypergeometrische kansverdeling

De kans op een bepaalde eigenschap is bij binomiale verdeling constant (of wordt modelmatig constant verondersteld). Daarom heet de binomiale kansverdeling ook een trekking *met teruglegging*. Het woord ‘terugleggen’ wordt hierbij letterlijk opgevat: het experiment wordt meerdere malen uitgevoerd, steeds onder de aanname dat de kans op ‘succes’ niet veranderd is en de trekkingen onafhankelijk van elkaar zijn. Bij de kansverdeling *zonder teruglegging* is de kans op ‘succes’ niet constant. Er worden uit een eindige populatie elementen gekozen, die na de trekking niet worden ‘teruggelegd’. Met andere woorden: de populatie verandert na elke trekking. De bijbehorende kansverdeling heet ook wel de *hypergeometrische kansverdeling*.

VOORBEELD 1.3

In een bedrijf moeten 60 posten worden gecontroleerd. Een accountant neemt een steekproef van 5 posten. Aangenomen wordt dat 12 van de 60 posten fout zijn. Hoe groot is de kans dat de accountant precies 3 foute posten in de steekproef aantreft? We nemen hierbij aan dat gecontroleerde posten op de stapel ‘al gecontroleerd’ terechtkomen, en dus niet ‘teruggelegd’ worden.

We berekenen de gevraagde kans op twee manieren. Stel, $\underline{k}$ is het aantal foute posten in de steekproef.

$$1 \quad P(\underline{k} = 3) = {}_5C_3 \cdot \frac{12}{60} \cdot \frac{11}{59} \cdot \frac{10}{58} \cdot \frac{48}{57} \cdot \frac{47}{56} = 0,0454$$

De kans dat de eerste post fout is, is 12/60. Omdat deze post op de stapel ‘al gecontroleerd’ komt, is de kans op een volgende foute post 11/59 en daarna 10/58. Daarna zijn er twee posten goed, dus met kansen 48/57 en 47/56. De vermenigvuldiging van deze vijf kansen (productregel) is de kans op de combinatie FFFGG (eerste 3 posten

fout, laatste 2 goed), vermenigvuldigd met het aantal mogelijke combinaties met 3 keer F en 2 keer G; dit is ${}_5C_3 = 10$.

$$2 \quad P(\underline{k} = 3) = \frac{{}_{12}C_3 \cdot {}_{48}C_2}{{}_{60}C_5} = \frac{220 \cdot 1128}{546512} = 0,0454$$

Bij deze methode wordt het aantal mogelijke combinaties met 3 fouten (uit 12) en 2 goede posten (uit 48) gedeeld door het totaal aantal mogelijke combinaties van 5 posten (uit 60).

De kans op 3 foute posten kan ook met Excel worden berekend, met de functie HYPGEOM.VERD(3;5;12;60;0), algemeen: HYPGEOM.VERD($k; n; M; N; 0/1$), met k het bestudeerde aantal 'gunstige' uitkomsten (hier: foute posten), n de steekproefgrootte, M het aantal elementen met bepaalde eigenschap, en N de populatiegrootte en als laatste de keuze voor de =-kans (0) of de cumulatieve $\leq$ -kans (1). Zie de pop-up in Excel hierna.

Functieargumenten

HYPGEOM.VERD

Steekproef-gunstig	3	= 3
Grootte-steekproef	5	= 5
Populatie-gunstig	12	= 12
Grootte-populatie	60	= 60
Cumulatief	0	= ONWAAR

= 0,045437967

Geeft als resultaat de hypergeometrische verdeling.

Cumulatief is een logische waarde: gebruik WAAR voor de cumulatieve verdelingsfunctie en gebruik ONWAAR voor de kansdichtheidsfunctie.

Resultaat formule = 0,045437967

[Help-informatie over deze functie](#)

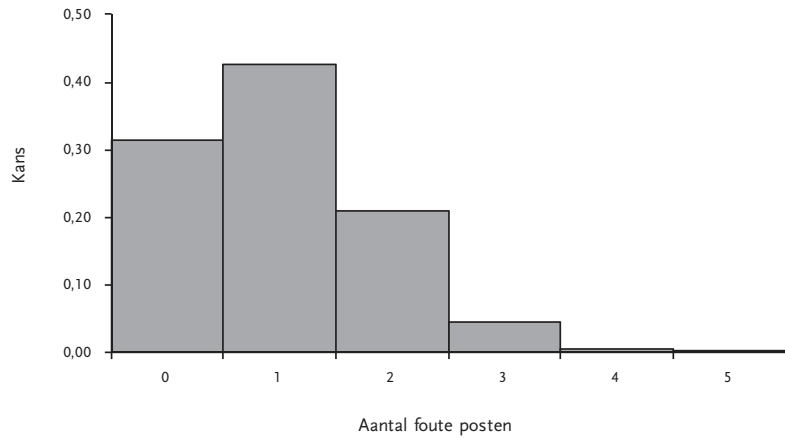
OK Annuleren

Vreemd genoeg bestaan er geen standaardfuncties voor de hypergeometrische kansverdeling bij de grafische rekenmachines. Natuurlijk kan men wel hypergeometrische kansen met de ${}_nC_r$ -functie berekenen.

Merk op dat de tweede methode niet wezenlijk verandert, als de waarde van k verandert. Voor bijvoorbeeld $k = 4$ (dus 4 foute posten en 1 goede) wordt de kans:

$$P(\underline{k} = 4) = \frac{{}_{12}C_4 \cdot {}_{48}C_1}{{}_{60}C_5} = \frac{495 \cdot 48}{546512} = 0,0044$$

Zo kan in principe de gehele kansverdeling worden bepaald. In figuur 1.2 is het histogram van deze hypergeometrische verdeling gegeven.



Figuur 1.2 Histogram van de hypergeometrische kansverdeling met $N = 60$, $M = 12$ en $n = 5$

De hypergeometrische verdeling wordt vaak aangegeven met de notatie $\underline{k} \sim \text{Hyp}(N; M; n)$. Theoretisch is de formule van een hypergeometrische kans dan:

$$P(\underline{k} = k) = \frac{{}^M C_k \cdot {}^{N-M} C_{n-k}}{{}^N C_n}$$

met:

- k het bestudeerde aantal ‘gunstige’ uitkomsten;
- n de steekproefgrootte;
- M het aantal elementen met een bepaalde eigenschap;
- N de populatiegrootte.

Op deze manier kan de kans op precies 3 foute posten formeel worden genoteerd als:

$$P(\underline{k} = 3 \mid \text{Hyp}(60; 12; 5)) = 0,0454$$

De verticale streep moet gelezen worden als ‘gegeven’ of ‘onder de voorwaarde dat’ (zie *Basisboek statistiek*, hoofdstuk 3, ‘Elementaire kansrekening’). Er zijn drie parameters nodig om de gehele hypergeometrische kansverdeling te kennen. Mede om deze reden wordt de hypergeometrische verdeling vaak benaderd door de binomiale kansverdeling, die maar twee parameters heeft. Een vuistregel zegt dat dit geoorloofd is als de steekproef kleiner is dan of gelijk is aan 10% van de populatiegrootte, ofwel als:

$$n \leq 0,1 \cdot N$$

De verwachtingswaarde (het ‘gemiddelde’) van de hypergeometrische verdeling kan gevonden worden door de verdeling te zien als benadering van de binomiale:

$$\mu = n \cdot p = n \cdot \frac{M}{N}$$

In voorbeeld 1.3 is dit gemiddelde aantal foute posten:

$$\mu = 5 \cdot \frac{12}{60} = 1$$

De standaardafwijking van de hypergeometrische kansverdeling wordt gegeven door:

$$\sigma = \sqrt{n \cdot \frac{M}{N} \cdot \frac{N-M}{N} \cdot \frac{N-n}{N-1}}$$

Deze nogal technische formule zullen we niet gebruiken in dit boek. In appendix 3 laten we zien dat deze standaardafwijking erg lijkt op de formule van de standaardafwijking van de binomiale verdeling. De factor:

$$\sqrt{\frac{N-n}{N-1}}$$

heet de *eindigepopulatiecorrectiefactor* en is nagenoeg gelijk aan 1, als $n \leq 0,1 \cdot N$. Deze correctiefactor voor eindige populaties zullen we ook nog terugzien bij de berekening van de steekproefgrootte in eindige populaties in paragraaf 2.6.

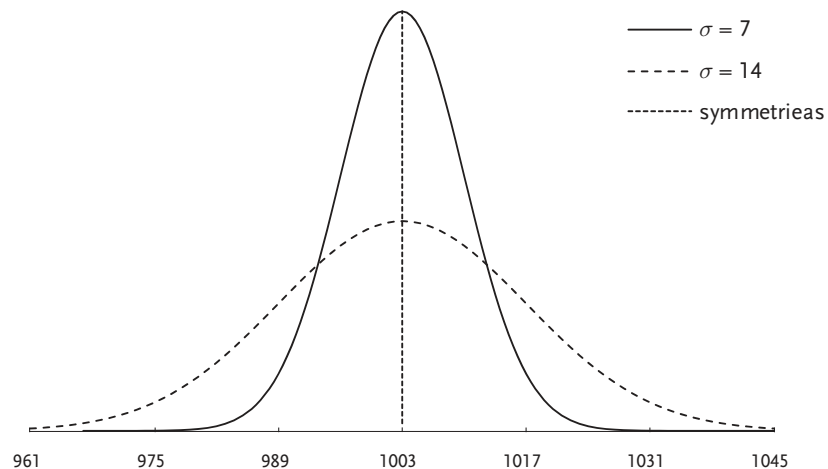
1.4 Normale kansverdeling

Naast de binomiale kansverdeling is er nog een kansverdeling die veel gebruikt wordt in de (economische) praktijk: de normale verdeling. De binomiale verdeling is een *discrete* kansverdeling; met alleen hele getallen als uitkomst. De kansgrafiek is een staafdiagram. De kansgrafiek van de normale kansverdeling (de ‘kromme van Gauss’) is een *continue* (‘gladde’) kromme; elk getal kan als uitkomst optreden. De vorm van een normale kansverdeling is die van een ‘klok’ of ‘hoed’ en de grafiek is symmetrisch. De twee karakteristieke parameters zijn het gemiddelde μ en de standaardafwijking σ . Zie voor een uitgebreide beschrijving hoofdstuk 6 in het *Basisboek statistiek*. Daarin wordt ook getoond dat de standaardafwijking ‘te zien’ is in een normale verdeling: het is de afstand van de symmetrieas tot een van beide buigpunten.

Een nauw aan de normale verdeling verwante kansverdeling, de zogeheten Student-*t*-verdeling, wordt in hoofdstuk 2 behandeld.

INTERMEZZO

De kennis over de normale verdeling is al ruim driehonderd jaar oud. De kansverdeling is onder verschillende namen gebruikt (onder andere als de kansverdeling van De Moivre, als de kansverdeling van Laplace, en als de errorkansverdeling). De Duitser Carl Friedrich Gauß (1777-1855) – ook wel geschreven als Gauss – was onder andere wiskundige, astronoom en landmeter. De afwijkingen van meetfouten die hij registreerde vormden bij benadering een kansverdeling die later de normale verdeling werd genoemd, bedoeld als ‘standaard’ kansverdeling.



Figuur 1.3 Voorbeelden van twee normale verdelingen: $\mu = 1003$ met $\sigma = 7$ en $\sigma = 14$

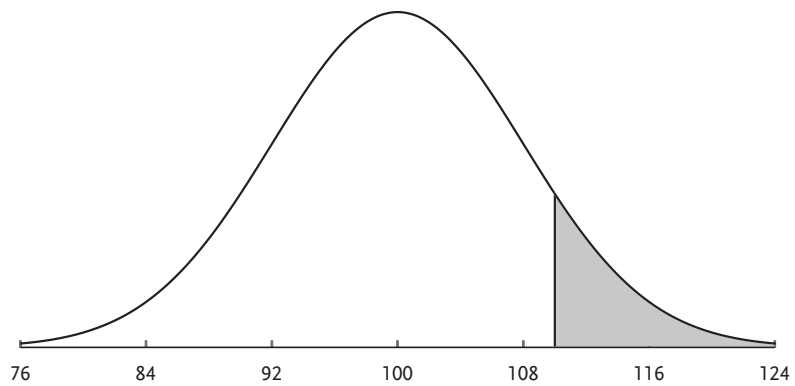
VOORBEELD 1.4

Stel dat de omzet van een supermarkt voorgesteld kan worden als een normaal verdeelde kansvariabele, met een gemiddelde (μ) van € 100.000 per week en een standaardafwijking (σ) van € 8000 per week.

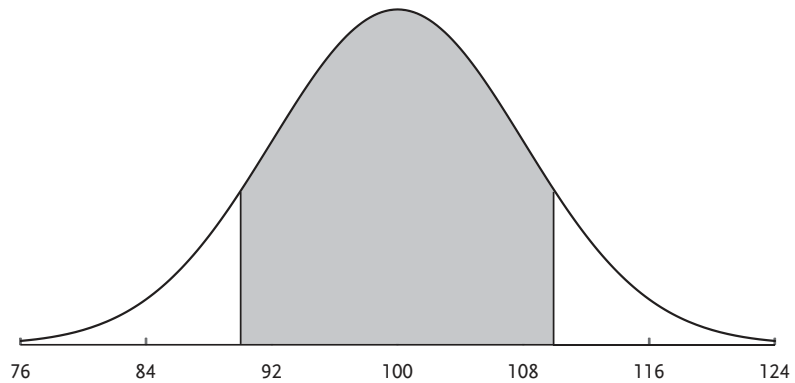
- 1 Hoe groot is de kans op een weekomzet van meer dan € 110.000?
- 2 Hoe groot is de kans op een weekomzet tussen € 90.000 en € 110.000?
- 3 Welke omzet X heeft de eigenschap dat de kans om X te overschrijden 75% is?

Oplossing:

- 1 Ter illustratie tekenen we de kansgrafiek van de omzet per week; zie figuur 1.4a. We nemen als eenheid € 1000 en gaan dus uit van $\mu = 100$ en $\sigma = 8$.



(a)



(b)

Figuur 1.4 Kans op een omzet (a) groter dan 110 en (b) tussen 90 en 110 bij een normale verdeling met $\mu = 100$ en $\sigma = 8$ ($\times \text{€ } 1000$)

We berekenen het antwoord (0,1056) op drie manieren:

Excel: $1 - \text{NORM.VERD.N}(110;100;8;1)$. (Oudere Excelversies dan 2010: NORM.VERD ; Engels: NORMDIST .) Excel kan alleen linkeroverschrijdingskansen berekenen, dus moet met behulp van de ontkenningregel gewerkt worden. De laatste 1 symboliseert (linkszijdig) cumulatief.