Poulen met EK-koorts

Miljoenen Nederlanders zullen in de aanloop naar Euro2008 meedoen met een EK-poule. Bij EK-poules gaat het erom dat men meer uitslagen juist weet te raden dan de andere deelnemers. Volgens de berekeningen staat Nederland in de kwartfinale.

Zoals bekend is het Nederlandse voetbalelftal ingedeeld in de zogenaamde poule des doods, samen met de WK-finalisten Italië en Frankrijk, en Roemenië, winnaar van dezelfde groep als Nederland tijdens de kwalificatie voorronde van het Europese kampioenschap. Hier zal worden ingegaan op het voorspellen van de voetbaluitslagen van de groepswedstrijden en welke teams zullen doorgaan naar de kwartfinales. Een voorspelling over wie de topscorer wordt, wie de finalisten worden en welk land kampioen zal worden blijft achterwege. De kwaliteit van de voorspellingen hangt af van de kwaliteit van de gebruikte data en de gebruikte methode.

Data

Op de site van de UEFA staat voor elk team op basis van in het (verre) verleden gespeelde europese wedstrijden een *competition record* (Tabel 1). Volgens deze lijst heeft bijvoorbeeld Nederland van de 117 wedstrijden er 72 gewonnen, 22 gelijkgespeeld en 23 verloren. Tellen we een gelijkspel als half winst, dan is het winpercentage voor Nederland gelijk aan 71 procent, zelfs hoger dan van de andere landen in deze poule. Nederland heeft in deze wedstrijden 241 keer gescoord en 91 doelpunten geïncasseerd, wat neerkomt op een scorend vermogen (μ) van 2,06 doelpunten per wedstrijd en een incasserend vermogen (ν) van 0,78 per wedstrijd. Ook deze kengetallen zijn beter dan van de anderen in Poule C, dus op grond van deze data is Nederland niet de underdog maar de favoriet in deze poule. Merk op dat bijna alle zestien landen, met uitzondering van Turkije, een hoger scorend dan incasserend vermogen hebben ($\mu/\nu>1$). De oorzaak hiervan is dat de landen die zich voor deze eindronde hebben gekwalificeerd overwegend sterke teams zijn. In de onderste rij valt af te lezen dat over alle 1478 wedstrijden in totaal 2605 doelpuntenvóór zijn gemaakt, ofwel 1,76 per wedstrijd tegenover 0,99 doelpunten tegen per wedstrijd. Het gemiddeld aantal doelpunten per wedstrijd, ofwel de *overall scoring context*, is dan 1,76 + 0,99 = 2,75. In de voorlaatste kolom (s) is de landenspecifieke scoring context gegeven, waarbij $s = \mu + \nu$. De overall scoring context voor een sport heeft een enorme impact op de *competitive balance*, de mate waarin teams aan elkaar gewaagd zijn (Groot, 2008). In basketbal vallen per wedstrijd gemiddeld meer dan honderd doelpunten, dus de scoring context is zeer hoog. Kleine verschillen in de sterkte tussen teams vertalen zich dan in grote verschillen in winpercentages en dus een lage competitive balance. Bij een lage scoring context daarentegen heeft het zwakkere team een veel grotere kans om niet te verliezen, bijvoorbeeld omdat de wedstrijd eindigt in 0–0, door een geluksdoelpunt of door arbitrale missers (Groot, 2004). Door de lage scoring context in het voetbal is het mogelijk dat een relatief zwak land als Griekenland Europees kampioen werd in 2004, door in de finalerondes drie keer op rij met 1–0 te winnen van de sterkere landen Frankrijk, Tsjechië en Portugal.

Methode

Hoewel er meerdere methoden zijn om voorspellingen op basis van geaggregeerde data te genereren worden hier de kansen op doelpunten geschat met behulp van de onafhankelijke Poissonverdeling (Kader 1). Kenmerkend voor deze verdeling is dat het optreden van een gebeurtenis, een doelpunt in een bepaald tijdsinterval, relatief zeldzaam, willekeurig en niet gerelateerd is aan voorgaande gebeurtenissen. Deze methode is uitermate geschikt om de kans op een bepaalde uitslag te voorspellen (Ryder, 2004). Stel dat men wil weten wat de beste voorspelling is voor de eerste wedstrijd van Nederland tegen Italië. De stochastische variabele van het aantal doelpunten gescoord door Nederland hangt dan niet alleen af van het scorend vermogen van Nederland, maar ook van het incasserend vermogen van Italië. Daarnaast is ook de scoring context voor deze wedstrijd van belang.

Resultaten

Op zowel de kwaliteit van de gebruikte data als de methode valt het een en ander af te dingen. Wat betreft de data zou het beter zijn geweest als er gegevens zouden zijn geweest over een groot aantal recent gespeelde wedstrijden tussen twee teams die in de poulefase tegen elkaar uitkomen, maar deze data zijn niet voorhanden omdat twee teams maar eens in de zoveel jaar tegen elkaar loten. Een andere dataset, bijvoorbeeld op basis van meer recente gegevens, genereert andere uitkomsten. Wat betreft de methode van de onafhankelijke Poissonverdeling is de voornaamste beperking dat als Italië tegen Nederland eerst scoort, dat dan de kans op een tweede doelpunt voor Italië of een tegendoelpunt van

LOEK GROOT
Universitair hoofddocent aan de Universiteit Utrecht.

Tabel 1

Historische gegevens van de deelnemende landen aan Euro2008.

	Aantal	Winst	Gelijk	Verlies	GF	GA	wpct	μ	ν	μ/ν	s	E
A												
ZWI	80	29	22	29	119	108	50%	1,49	1,35	1,10	2,84	3,1
TSJ	54	39	8	7	117	37	80%	2,17	0,69	3,16	2,85	6,5
POR	108	58	26	24	183	98	66%	1,69	0,91	1,87	2,60	4,9
TUR	95	35	22	38	110	135	48%	1,16	1,42	0,81	2,58	2,2
B												
OOS	80	33	13	34	146	125	49%	1,83	1,56	1,17	3,39	2,7
KRO	47	28	11	8	86	38	71%	1,83	0,81	2,26	2,64	5,0
DUI	110	67	29	14	224	77	74%	2,04	0,70	2,91	2,74	5,8
POL	90	38	24	28	131	100	56%	1,46	1,11	1,31	2,57	3,1
C												
NED	117	72	22	23	241	91	71%	2,06	0,78	2,65	2,84	4,7
ROE	105	52	27	26	191	104	62%	1,82	0,99	1,84	2,81	3,4
FRA	117	67	29	21	236	109	70%	2,02	0,93	2,17	2,95	4,0
ITA	111	59	35	17	175	77	69%	1,58	0,69	2,27	2,27	4,2
D												
GRI	100	48	19	33	146	115	58%	1,46	1,15	1,27	2,61	3,0
ZWE	93	44	24	25	140	91	60%	1,51	0,98	1,54	2,48	3,6
SPA	123	72	24	27	260	111	68%	2,11	0,90	2,34	3,02	5,1
RUS	48	27	10	11	100	48	67%	2,08	1,00	2,08	3,08	4,7
Totaal	1478				2605	1464	64%	1,76	0,99			

Nederland onafhankelijk is van het eerste doelpunt. Met andere woorden, het wedstrijdverloop doet er niet toe, terwijl een 1–0-voorsprong voor Italië zelden uit handen wordt gegeven. Ook het thuisvoordeel voor Zwitserland en Oostenrijk is niet meegenomen. Gegeven deze beperkingen worden hieronder de volgende voorspellingen gegenereerd: per wedstrijd in de poulefase de kansen op winst, verlies en gelijkspel, de meest waarschijnlijke uitslagen met de bijbehorende kansen, Het verwachte aantal doelpunten voor en tegen, en de teams die zich zullen kwalificeren voor de kwartfinales.

In tabel 2 zijn op basis van de scoringsintensiteiten μ en ν voor elk land en de landenspecifieke scoringcontextvariabele s uit tabel 1 de wedstrijdspecifieke Poissonparameters μ* en ν* berekend. Voor de wedstrijd Nederland–Italië is de parameter μ* berekend als het gemiddeld scorend vermogen van Nederland (2,06) maal het gemiddeld incasserend vermogen van Italië (0,69) gedeeld door de helft van de gemiddelde landenspecifieke scoring context 0,5×((2,84 + 2,27)/2). We nemen de helft van de gemiddelde scoring context omdat de nieuwe parameters samen weer bij benadering gelijk moeten zijn aan de gemiddelde scoring context. Het scorend vermogen van Nederland in deze wedstrijd (μ* = 1,12) is wedstrijdspecifiek omdat de waarde afhangt van de geaggregeerde gegevens voor beide landen, maar niet van de andere landen. Merk op dat het scorend en incasserend vermogen van Nederland verschilt per wedstrijd: tegen Italië is het lager dan tegen Frankrijk en Roemenië omdat het incasserend vermogen van Italië lager is. Analoog wordt het scorend vermogen van Italië (ν* = 0,96) voor deze wedstrijd bepaald door het scorend vermogen van Italië (1,58), het incasserend vermogen van Nederland (0,78) en de scoring context voor deze wedstrijd. Met dit paar parameters kan de kans op een bepaalde uitslag worden berekend. Bijvoorbeeld, de kans op de uitslag 2–1 is gelijk aan het product van P(N=2) en P(I=1). Voor de bepaling van de kansen zijn alleen de uitslagen met per wedstrijd maximaal zeven doelpunten of minder meegenomen, omdat de kansen op bijvoorbeeld 8–0, 5–3 of 4–4 verwaarloosbaar klein zijn.

Bij de kolommen Winst, Verlies en Gelijkspel wordt bijvoorbeeld de kans op een gelijkspel berekend door de kansen op de uitslagen 0–0, 1–1, 2–2 en 3–3 te sommeren. De kolom U* geeft de meest waarschijnlijke uitslagen en P* de bijbehorende kansen. Wat opvalt aan deze lijst, zeker als iemand dit zou invullen in een poule, is dat ze nogal aan de eentonige kant is. Van de 24 wedstrijden eindigen er maar liefst elf in 1–0, nog eens tien in 1–1 en drie in 2–0. In totaal vallen er volgens deze lijst slechts 37 doelpunten in 24 wedstrijden, ofwel gemiddeld 1,5 doelpunt per wedstrijd, wat veel minder is dan de overall scoring context. De oorzaak hiervan is dat naast de meest waarschijnlijke uitslag er nog een hele waaier minder waarschijnlijke uitslagen zijn met veelal meer doelpunten maar elk met een lagere kans. Voor Nederland–Italië is 1–0 de meest waarschijnlijke uitslag met een kans van 14,0 procent en alleen de uitslag 0–0 met 12,5 procent heeft minder doelpunten; alle andere uitslagen zoals 1–1 (13,4 procent), 2–0 (7,8 procent), 3–0 (2,9 procent) et cetera hebben meer doelpunten. Dat deelnemers aan poules doorgaans kiezen voor een wat feestelijker lijst van minder waarschijnlijke uitslagen met meer doelpunten kan mogelijk hieruit worden verklaard dat men dan tijdens de wedstrijd langer in de race blijft. Bij een voorspelde uitslag van 0–0 of 1–0 is het al gedaan als Italië scoort, maar bij 2–2 blijft tot het einde toe de kans open dat het uitkomt. In de sporteconomie staat dit bekend als de *longshot bias*. Een andere reden is dat meer extreme uitslagen de kans verhogen dat de pot niet gedeeld hoeft te worden met andere winnaars.

De laatste twee kolommen van tabel 2 geven per wedstrijd de verwachtingswaarde van het aantal gescoorde doelpunten voor en tegen (GF* en GA*). Deze wordt verkregen wordt verkregen door per wedstrijd alle (36, want gemakshalve beperkt tot maximaal zeven doelpunten per wedstrijd) mogelijke uitkomsten te wegen met de kansen. Voor de wedstrijd Nederland–Italië zijn de verwachtingswaarden 1,11 doelpunt voor Nederland en 0,95 doelpunt voor Italië. Afgerond

Kader 1

In de Poisson verdeling is de kans dat een team in een wedstrijd N doelpunten maakt gelijk aan: $P(N) = \frac{e^{-\mu}\mu^N}{N!}$

met e het grondtal van de natuurlijke logaritme (e = 2,71828), N! de faculteit van N (N*(N-I)*...(I)) en μ het gemiddeld aantal gescoorde doelpunten per wedstrijd voor dat team. Als twee teams tegen elkaar uitkomen, dan zijn voor de voorspelling van de uitslag twee parameters nodig: het scorend vermogen μ van het ene team (per definitie gelijk aan het incasserend vermogen van het andere team) en het scorend vermogen van de opponent ν (idem). De kans op de uitslag N-I is dan:

$$P(N\text{-}I) = P(N)^* P(I) = \frac{e^{-\mu}\mu^N}{N!} * \frac{e^{-\nu}\nu^I}{I!}$$

De kans op een gelijkspel is de som van alle kansen waar N = I, de kans op winst de som waarbij N>I en de kans op verlies de som waarbij N<I.

Tabel 2

Berekening van de wedstrijdspecifieke scoringparameters, de kansen op winst, verlies en gelijkspel, de meest waarschijnlijke uitslag en verwacht aantal doelpunten voor en tegen per wedstrijd.

		μ*	ν*	Winst	Verlies	Gelijk	U*	P*	GF*	GA*
A										
ZWI	TSJ	0,72	2,06	12%	69%	19%	0,2	13%	0,70	2,01
POR	TUR	1,86	0,81	62%	16%	22%	1,0	13%	1,82	0,80
TSJ	POR	1,44	0,85	51%	22%	27%	1,0	15%	1,43	0,84
ZWI	TUR	1,56	1,15	47%	28%	25%	1,1	12%	1,53	1,13
ZWI	POR	0,99	1,68	22%	54%	24%	0,1	12%	0,97	1,65
TUR	TSJ	0,58	2,27	8%	76%	16%	0,2	15%	0,57	2,21
B										
OOS	KRO	0,98	1,90	19%	59%	22%	0,1	11%	0,95	1,85
DUI	POL	1,71	0,77	60%	17%	24%	1,0	14%	1,68	0,76
KRO	DUI	0,95	1,23	29%	42%	29%	0,1	14%	0,95	1,22
OOS	POL	1,36	1,53	34%	41%	25%	1,1	12%	1,32	1,48
POL	KRO	0,90	1,56	22%	53%	25%	0,1	13%	0,89	1,54
OOS	DUI	0,83	2,08	14%	66%	20%	0,2	12%	0,81	2,02
C										
ROE	FRA	1,18	1,39	32%	42%	26%	1,1	13%	1,16	1,37
NED	ITA	1,12	0,96	39%	31%	30%	1,0	14%	1,11	0,95
ITA	ROE	1,23	0,99	42%	30%	29%	1,0	13%	1,22	0,99
NED	FRA	1,33	1,08	42%	30%	27%	1,1	13%	1,31	1,07
NED	ROE	1,45	1,00	47%	26%	27%	1,1	13%	1,43	0,99
FRA	ITA	1,07	1,13	34%	37%	29%	1,1	13%	1,06	1,12
D										
SPA	RUS	1,39	1,23	40%	33%	26%	1,1	12%	1,36	1,21
GRI	ZWE	1,12	1,36	31%	42%	27%	1,1	13%	1,11	1,34
ZWE	SPA	0,99	1,50	25%	49%	26%	0,1	12%	0,97	1,48
GRI	RUS	1,03	1,68	23%	53%	24%	1,1	11%	1,00	1,65
GRI	SPA	0,94	1,73	20%	56%	24%	0,1	12%	0,92	1,69
RUS	ZWE	1,46	1,08	46%	28%	26%	1,1	12%	1,44	1,06

is de uitslag op basis van de verwachtingswaarden 1–1, maar deze uitslag heeft een lagere kans (13,4 procent) dan de meest waarschijnlijke uitslag 1–0 (14,0 procent). Als men in de poule maar één uitslag mag invullen, is toch de eerste te verkiezen boven de laatste. De verwachtingswaarden kunnen echter worden gebruikt voor weddenschappen bij bookmakers waarbij men het aantal doelpunten in een wedstrijd moet voorspellen, of het aantal doelpunten voor of tegen per team. Merk op dat het totaal aantal volgens de verwachtingswaarden gescoorde doelpunten (60,2) veel hoger ligt dan volgens de lijst meest waarschijnlijke uitslagen (37).

In de laatste kolom van tabel 1 is voor elk land het verwachte aantal wedstrijdpunten gegeven, waarbij winst met drie en een gelijkspel met één punt is gewaardeerd. Poule C is veruit de spannendste (de standaardafwijking in het puntentotaal is slechts 0,4) terwijl poule A op voorhand als het minst spannend kan worden aangemerkt (standaardafwijking 1,6). Volgens deze berekeningen en het wedstrijdschema na de poulefase, waarbij de winnaar van poule A moet spelen tegen de runner-up van poule B, zijn de kwartfinales Tsjechië–Kroatië, Duitsland–Portugal, Nederland–Rusland en Spanje–Italië.

Naar de bookmakers

Bookmakers zoals BetExplorer en Unibet publiceren zogenaamde *odd ratios*. Bij BetExplorer zijn voor de wedstrijd Nederland–Italië de *odds* op Winst, Verlies en Gelijkspel gelijk aan 2,90, 2,40 en 3,06. Als men een euro inzet op winst dan wordt 2,9 euro uitgekeerd als Nederland inderdaad wint. Volgens de berekeningen in tabel 2 zijn de kansen op Winst, Verlies en Gelijkspel gelijk aan 39,1 procent, 30,9 procent en 30,0 procent, samen 100 procent. De actuariële *fair odds* kunnen worden berekend door de reciproque te nemen, dus 1/0,391 = 2,56, 1/0,309 = 3,23 en 1/0,300 = 3,33. Omdat de bookmakers winst moeten maken zijn hun odds niet fair, want de som van hun impliciete kansen (1/2,90 + 1/2,40 + 1/3,06) = (0,345 + 0,417 + 0,327) = 108,9 procent. Het surplus van 8,9 procent wordt wel de *overround* of de *vig* genoemd en is een belangrijk bestanddeel van de winst voor de bookmaker (naast het exploiteren van biases onder deelnemers, zoals de longshot bias). Ondanks de overround is het soms mogelijk dat een bepaalde odd van de bookmaker hoger is dan de fair odd, ofwel de bookmaker schat de kans lager in dan de feitelijke kans en keert ten gevolge daarvan te veel uit bij het optreden van de gebeurtenis (Groot, 2007). Voor Nederland–Italië is dit het geval voor Winst: gegeven een feitelijke kans van 39,1 procent zou per ingelegde euro de faire uitkering 2,56 euro bedragen, maar de bookmaker keert 2,90 euro uit. Per ingelegde euro is de verwachte winst 0,391 × 2,9 – 1= 0,13 ofwel 13 cent. Hierbij moet wel worden bedacht dat bookmakers waarschijnlijk verschillende, en meer geavanceerde, methoden combineren, om de systematische fouten van de ene methode uit te middelen met die van andere, om de odds te bepalen. Daarnaast hebben ze de beschikking over meer accurate data en zullen experts indien nodig de parameters bijstellen op grond van fingerspitzengefühl. Het op het laatste moment afhaken van Clarence Seedorf voor het EK kan tot gevolg hebben dat de parameters van Nederland worden bijgesteld. Niettemin komen twintig van de 24 uitslagen overeen met de laagst genoteerde odds bij bookmaker Ladbrokes. Ten slotte dekken bookmakers zich in tegen deelnemers die systematisch de bookmaker proberen te verslaan door bepalingen op te nemen van maximaal uit te keren bedragen.

Conclusie

Het voorspellen van voetbaluitslagen blijft koffiedik kijken. De kans dat alle hier voorspelde uitslagen fout zijn is vele malen groter dan de kans dat alle goed zijn. Bij EK-poules gaat het echter niet om alle dertien goed, zoals bij de Toto, maar dat men meer uitslagen juist weet te raden dan de andere deelnemers. Ondanks alle beperkingen is de kans zeer reëel dat enkele uitslagen goed zijn. Naast de hierboven genoemde beperkingen produceren ook de scheidsrechters veel ruis door het nemen van foutieve maar cruciale beslissingen. Ook de laatste pouleduels kunnen om strategische redenen, zoals het aansturen op een 0–0, anders worden gespeeld dan in de Poissonverdeling voor doelpunten is verondersteld.

LITERATUUR

Groot, L.F.M. (2004) Scheidsrechter is de belangrijkste speler. *NRC Handelsblad*, 22–06–2004, Opinie-pagina 7.

Groot, L.F.M. (2007) *Economics, uncertainty and European football: trends in competitive balance.* Cheltenham en Northampton MA: Edward Elgar.

Groot, L.F.M. (2008) Some determinants of the natural level of competitive balance in European football and US team sports: the role of the referee, the scoring context and overtime. Paper gepresenteerd bij het 10de IASE congres, Gijon, 9–10 mei 2008.

Ryder, A. (2004) *Poisson Toolbox: A Review of the Application of the Poisson Probability Distribution in Hockey*, www.HockeyAnalytics.com.

UEFA website, http://en.euro2008.uefa.com/tournament/statistics/index.html.