Simuleren van Verwachtingswaarden

Hans van der Weide*

18 januari 2000

Simulatie kan in het VWO onderwijs een belangrijke rol spelen bij het illustreren en toelichten van verschijnselen waarbij het toeval een belangrijke rol speelt. We zullen hier enige voorbeelden van simulatie behandelen aan de hand van het pakket WinStats. Het gaat om voorbeelden waarbij de verwachtingswaarde van een stochast gesimuleerd wordt. Van de voorbeelden die wij zullen bespreken kan de verwachtingswaarde ook exact berekend worden zodat het mogelijk is om de resultaten van de simulatie te beoordelen. Verder geven we achtergrondinformatie over de theorie waarop het simuleren gebaseerd is.

1 Simuleren van kansverdelingen

Iemand die met een dobbelsteen gaat gooien weet dat de dobbelsteen op één van de getallen 1, 2, 3, 4, 5 of 6 zal vallen. Als hij, zoals de meesten van ons, niet over bovennatuurlijke gaven beschikt en ook niet in staat is om bij het gooien de uitkomst te manipuleren, dan is dit alles dat gezegd kan worden over het resultaat van één worp met een dobbelsteen: de uitkomst van één worp met een dobbelsteen is onvoorspelbaar.
Als we nu een groot aantal malen met de dobbelsteen gooien dan weten we uit ervaring dat bij een zuiver geconstrueerde dobbelsteen elk van de uitkomsten ongeveer even vaak zal voorkomen. Men duidt dit verschijnsel aan als de **Empirische Wet van de Grote Aantallen**. In een wiskundig model voor een worp met een zuivere dobbelsteen kennen we getallen $p_1, \ldots, p_6$ toe aan de uitkomsten. Deze getallen corresponderen met de relatieve frequenties van de uitkomsten. Let op, de getallen p_1 tot en met p_6 worden door ons gekozen. Voor een zuivere dobbelsteen kiezen we $p_1 = \ldots = p_6 = \frac{1}{6}$.
Het gooien van een dobbelsteen kunnen we imiteren op een computer door gebruik te maken van een random number generator.[1] Een random generator is een computerprogramma dat getallen produceert die beschouwd mogen worden als aselecte trekkingen uit de getallen tussen 0 en 1. Een voorbeeld van zo'n programma is de multiplicatieve congruentie generator. Zij a een geheel getal en m een positief geheel getal, dan is $a(\mathrm{mod} m)$ de rest bij deling van a door m. Om een rij random getallen $X_0, X_1, X_2, \ldots$ te genereren kiezen we een geheel getal als startwaarde X_0. Voor $n = 1, 2, \ldots$ is dan

$$X_{n+1} = (aX_n + c)(\mathrm{mod} m),$$

waarbij a, c en m (zorgvuldig) gekozen constanten zijn. De aldus gevormde rij $X_0, X_1, X_2, \ldots$ is een rij gehele getallen tussen 0 en $m-1$. Om een rij reële getallen

*TU Delft, faculteit Informatietechnologie en Systemen

[1]Zie voor meer informatie het boek INTRODUCTION TO PROBABILITY geschreven door Charles M. Grinstead en J. Laurie Snell. Dit boek is vrij toegankelijk via de website: http://www.dartmouth.edu/~ chance/teaching_ aids/books_ articles/probability_ book/pdf.html

tussen 0 en 1 te krijgen delen we iedere X_j door m. in Als we met een computer het random getal u genereren, dan is $1 + \lfloor 6u \rfloor$ een uitkomst van een worp met een zuivere dobbelsteen. Door deze procedure te herhalen kunnen we het herhaald gooien met een dobbelsteen op de computer naspelen. Het spreekt vanzelf dat er bepaalde eisen gesteld moeten worden aan een random number generator om te kunnen rechtvaardigen dat de resultaten van het computer experiment overeenkomen met het herhaald gooien van een echte dobbelsteen. Hiervoor bestaat een groot aantal tests. Een waarschuwing is hier op zijn plaats, zelfs in professionele pakketten zijn nog wel eens zeer slechte random number generatoren ingebouwd.
We onderbreken dit betoog kort om een paar relevante begrippen uit de kansrekening op te halen. Een stochast geeft een getal bij iedere uitkomst van het kansexperiment. Bij het gooien met een dobbelsteen is de uitkomst zelf een stochast. Een ander voorbeeld is de stochast X die gelijk is aan het aantal I's als de uitkomst als Romeins cijfer geschreven wordt. Dan is $\mathbf{P}(X = 0) = 1/6$, $\mathbf{P}(X = 1) = 3/6$, $\mathbf{P}(X = 2) = 1/6$ en $\mathbf{P}(X = 3) = 1/6$. Dit kunnen we ook op de volgende manier afkorten:

$$\begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{1}{6} & \frac{3}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}.$$

Stel dat we n-maal gooien met de dobbelsteen en dat daarbij de frequentie van de gebeurtenis $\{X = i\}$ gelijk is aan n_i, $i = 0, 1, 2, 3$. Dus $n_0 + n_1 + n_2 + n_3 = n$ en het gemiddeld aantal I's is gelijk aan:

$$\frac{0 \cdot n_0 + 1 \cdot n_1 + 2 \cdot n_2 + 3 \cdot n_3}{n} = 0 \cdot \frac{n_0}{n} + 1 \cdot \frac{n_1}{n} + 2 \cdot \frac{n_2}{n} + 3 \cdot \frac{n_3}{n}.$$

In het kansmodel definiëren we de verwachtingswaarde van X door:

$$\mathbf{E}(X) = 0 \cdot \frac{1}{6} + 1 \cdot \frac{3}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} = \frac{4}{3}.$$

Uit de opzet van het kansmodel zien we dat voor grote waarden van n het gemiddeld aantal I's bij het kansexperiment ongeveer gelijk zal zijn aan de verwachtingswaarde van X.
We eindigen deze paragraaf met de definitie van de variantie van een stochast Y. Stel dat de verwachtingswaarde van Y gelijk is aan $\mathbf{E}(Y) = m$, dan wordt de variantie d^2 van Y gegeven door:

$$\begin{aligned} d^2 = \mathbf{E}\left((Y - m)^2\right) &= \mathbf{E}\left(Y^2 - 2mY + m^2\right) \\ &= \mathbf{E}(Y^2) - 2m\mathbf{E}(Y) + m^2 = \mathbf{E}(Y^2) - m^2, \end{aligned}$$

waarbij we gebruik gemaakt hebben van de somregel die zegt dat de verwachting van een som gelijk is aan de som van de verwachtingen. Bijvoorbeeld: de variantie van X is gelijk aan:

$$\mathrm{Var}(X) = \frac{16}{6} - \left(\frac{4}{3}\right)^2 = \frac{8}{9}.$$

2 Wet van de Grote Aantallen

In de vorige paragraaf hebben we gezien dat de gemiddelde waarde van een stochast bij een groot aantal herhalingen van een kansexperiment in de buurt zal liggen van de verwachtingswaarde van deze stochast. In deze paragraaf willen ingaan op wat meer achtergronden. Beschouw onafhankelijke stochasten $X_1, X_2, \ldots$ met dezelfde verdeling met verwachting m en variantie d^2. Zij verder $\bar{X}_n =$

$(X_1 + \cdots + X_n)/n$ het gemiddelde van X_1 tot en met X_n. Dan geldt voor de verwachtingswaarde

$$\mathbf{E}(\bar{X}_n) = m.$$

De kansverdeling van het gemiddelde is dus geconcentreerd rond m, maar we weten nog niet 'hoe dicht' $\bar{X}_n$ ligt bij m. Bekijk hiervoor de volgende kans

$$\mathbf{P}(|\bar{X}_n - m| > \epsilon)$$

voor een (klein) getal $\epsilon > 0$. De grootte van deze kans voor verschillende keuzes van ϵ zegt iets over 'hoe dicht' $\bar{X}_n$ bij m ligt. Zolang we niets weten over de kansverdeling van de X_i's kunnen we deze kans niet uitrekenen, we kunnen wel een algemeen geldende bovengrens vinden.

Lemma 2.1 *(***Ongelijkheid van Tchebycheff***)*
Zij Y een stochast met $\mathbf{E}(Y) = m$ en zij $\epsilon > 0$. Dan geldt:

$$\mathbf{P}(|Y - m| > \epsilon) \leq \frac{\mathrm{Var}(Y)}{\epsilon^2}. \tag{1}$$

Bewijs: Definieer de stochast Z door het voorschrift:

$$Z = \begin{cases} \epsilon & , \text{ als } |Y - m| > \epsilon \\ 0 & , \text{ als } |Y - m| \leq \epsilon. \end{cases}$$

Dan is Z een discrete stochast met mogelijke waarden 0 en ϵ en kansen:

$$\begin{aligned} \mathbf{P}(Z = 0) &= \mathbf{P}(|Y - m| \leq \epsilon) \\ \mathbf{P}(Z = \epsilon) &= \mathbf{P}(|Y - m| > \epsilon). \end{aligned}$$

Merk op dat per definitie $0 \leq Z \leq |Y - m|$, dus ook $Z^2 \leq (Y - m)^2$. Hieruit volgt:

$$\mathbf{E}(Z^2) \leq \mathbf{E}\left((Y - m)^2\right) = \mathrm{Var}(Y).$$

Maar

$$\mathbf{E}(Z^2) = 0^2 \cdot \mathbf{P}(Z = 0) + \epsilon^2 \cdot \mathbf{P}(Z = \epsilon) = \epsilon^2 \mathbf{P}(|Y - m| > \epsilon).$$

Zodoende is

$$\epsilon^2 \mathbf{P}(|Y - m| > \epsilon) \leq \mathrm{Var}(Y).$$

Delen door ϵ^2 levert de ongelijkheid van Tchebycheff. ■

Beschouw weer onafhankelijke stochasten $X_1, X_2, \ldots$ met dezelfde verdeling met verwachting m en variantie d^2. Omdat $\mathbf{E}(\bar{X}_n) = m$ levert toepassing van de ongelijkheid van Tchebycheff op $Y = \bar{X}_n$:

$$\mathbf{P}(|\bar{X}_n - m| > \epsilon) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon^2}.$$

Omdat de variantie van het gemiddelde wordt gegeven door

$$\mathrm{var}(\bar{X}_n) = \frac{d^2}{n},$$

vinden we dat

$$\mathbf{P}(|\bar{X}_n - m| > \epsilon) \leq \frac{d^2}{n\epsilon^2}.$$

Hieruit volgt dat

$$\lim_{n \to \infty} \mathbf{P}(| \bar{X}_n - m |> \epsilon) = 0. \tag{2}$$

Met andere woorden, de kans dat $\bar{X}_n$ en m meer dan een gegeven tolerantie ϵ van elkaar verschillen is willekeurig klein te maken door n voldoende groot te kiezen. Men noemt eigenschap (2) de **Zwakke Wet van de Grote Aantallen**. In feite geldt een sterkere uitspraak:

$$\mathbf{P}(\lim_{n\to\infty} \bar{X}_n = m) = 1. \tag{3}$$

Eigenschap (3) is de **Sterke Wet van de Grote Aantallen**. We gaan hier niet in op het precieze verschil tussen de Zwakke en de Sterke Wet van de Grote Aantallen.

Voorbeeld. Definieer de stochast X_i gelijk aan 1 als bij het i-de experiment gebeurtenis A optreedt en anders gelijk aan 0, en zij p_A de kans op A. In ons model voor herhalingen van experimenten zijn de stochasten X_i onafhankelijk en gelijk verdeeld met verdeling:

$$X_i : \begin{pmatrix} 0 & 1 \\ 1-p_A & p_A \end{pmatrix}.$$

De verwachting m van elke X_i is dan gelijk aan

$$m = 0 \cdot (1-p_A) + 1 \cdot p_A = p_A.$$

Verder is in dit geval $\bar{X}_n$ gelijk aan de relatieve frekwentie van de gebeurtenis A:

$$\bar{X}_n = \frac{N_n(A)}{n}.$$

Uit de Sterke Wet van de Grote Aantallen volgt dat

$$\lim_{n\to\infty} \frac{N_n(A)}{n} = p_A \quad \text{met kans 1.}$$

Dit bevestigt onze veronderstelling dat de relatieve frekwentie $N_n(A)/n$ voor grote waarden van n ongeveer gelijk is aan de kans op A.

3 Roll dice...

Een van de simulatie experimenten in het programma *WinStats* is het D-maal gooien van een N-dobbelsteen, dit is een zuivere dobbelsteen met N kanten. Zij X het aantal verschillende kanten van de dobbelsteen dat bij een uitvoering van dit experiment bovenkomt. Het berekenen van de kansverdeling van de stochast X vraagt enige slimme combinatoriek. Daarom willen we met behulp van simulatie onderzoeken wat de verwachtingswaarde $\mathbf{E}(X) = \mu_{N,D}$ is van het aantal verschillende kanten dat bovenkomt bij D worpen met de dobbelsteen. Hiertoe herhalen we dit experiment $n = 1000$ maal. Zij X_k het aantal verschillende kanten dat bovenkomt bij de k^{de} uitvoering van dit experiment. Uit de Sterke Wet van de Grote Aantallen volgt dat

$$\frac{1}{1000}\sum_{k=1}^{1000} X_k \approx \mu_{N,D}.$$

In dit geval kunnen we de verwachtingswaarde $\mu_{N,D}$ ook berekenen zonder eerst de kansverdeling van X uit te rekenen. Definieer voor $n = 1, \ldots, D$ de stochast T_n als het aantal verschillende kanten na n worpen met de dobbelsteen. Als gegeven is dat $T_{n-1} = i$ dan kan T_n alleen maar gelijk zijn aan i of aan $i+1$: $T_n = i$ als bij de n^{de} worp een van de eerder gegooide kanten bovenkomt en anders is $T_n = i + 1$. Dus:

$$\mathbf{P}(T_n = i \mid T_{n-1} = i) = \frac{i}{N} \quad \text{en} \quad \mathbf{P}(T_n = i+1 \mid T_{n-1} = i) = \frac{N-i}{N}.$$

Hieruit volgt dat
$$\mathbf{E}(T_n \mid T_{n-1} = i) = \frac{N-1}{N} i + 1.$$
Dus
$$\begin{aligned}
\mathbf{E}(T_n) &= \sum_{i=1}^{N} \mathbf{E}(T_n \mid T_{n-1} = i)\mathbf{P}(T_{n-1} = i) \\
&= \sum_{i=1}^{N} \left(\frac{N-1}{N} i + 1\right) \mathbf{P}(T_{n-1} = i) \\
&= \frac{N-1}{N} \mathbf{E}(T_{n-1}) + 1.
\end{aligned}$$
De oplossing van deze 1^e orde lineaire differentievergelijking is van de vorm
$$\mathbf{E}(T_n) = a \left(\frac{N-1}{N}\right)^n + b.$$
Substitutie in de differentievergelijking geeft $b = N$. Men gaat eenvoudig na dat $\mathbf{E}(T_1) = 1$, dus $a = -N$ en
$$\mathbf{E}(T_n) = -N \left(\frac{N-1}{N}\right)^n + N, \quad n = 1, \dots, D.$$
Dus:
$$\mu_{N,D} \;=\; \mathbf{E}(T_D) = N \left(1 - \left(\frac{N-1}{N}\right)^D\right).$$

4 De Centrale Limietstelling

In deze paragraaf gaan we nader op het random gedrag van de afwijkingen van het gemiddelde en de verwachtingswaarde zoals besproken in paragraaf 2. Als we de kansverdeling van het gemiddelde
$$\frac{X_1 + X_2 + \cdots + X_n}{n}$$
van onafhankelijke stochasten met dezelfde verdeling bekijken, dan blijkt dat deze kansverdeling voor $n \to \infty$ steeds meer gaat lijken op één en dezelfde limietkansverdeling, ongeacht de verdeling van de X_i's (als de variantie maar eindig is). Dit resultaat wordt precies gedefinieerd in de nu volgende Stelling.

Stelling 4.1 *(***Centrale Limietstelling***)*
Laat $X_1, X_2, \dots$ onafhankelijke stochasten zijn met dezelfde verdeling met verwachting m en variantie d^2. Zij verder $\bar{X}_n = (X_1 + \cdots + X_n)/n$ het gemiddelde van X_1 tot en met X_n. Dan convergeert de verdelingsfunctie van
$$\sqrt{n}\frac{\bar{X}_n - m}{d}$$
naar de verdelingsfunctie van de standaard normale verdeling. Anders gezegd:
$$\mathbf{P}\left(\sqrt{n}\frac{\bar{X}_n - m}{d} \le x\right) \to \Phi(x)$$
voor iedere x, waarbij
$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\, du,$$
de verdelingsfunctie is van de standaard normale verdeling.

De Centrale Limietstelling zegt dus onder andere dat we het verschil $\bar{X}_n - m$ moeten opblazen met een factor $\sqrt{n}$ om in de limiet iets over te houden. Bovendien zegt de stelling dat we in de limiet altijd hetzelfde overhouden, namelijk na deling door d houden we altijd een standaard normale verdeling over. Dit laatste geeft dus aan hoe realisaties van $\bar{X}_n$ rond m ongeveer fluctueren. De praktische toepassing van de Centrale Limietstelling is dat we kansen voor het gemiddelde van n onafhankelijke stochasten kunnen benaderen door overeenkomstige kansen van de standaard normale verdeling, zolang n vrij groot is:

$$\mathbf{P}\left(\sqrt{n}\frac{\bar{X}_n - m}{d} \leq x\right) \approx \mathbf{P}(Z \leq x)$$

waarbij Z standaard normaal verdeeld is.

4.1 Toepassingen van de Centrale Limietstelling

We zullen aan de hand van twee voorbeelden toepassingen van de Centrale Limietstelling illustreren. Het eerste voorbeeld is een directe toepassing van de stelling. Het tweede voorbeeld illustreert hoe kansen van een binomiale verdeling kunnen worden benaderd door kansen van de standaard normale verdeling.
Stel dat we 100 keer een willekeurig getal willen trekken tussen 0 en 1. Wat is de kans dat de som van de getrokken getallen ≥ 55 is? We maken het volgende kansmodel voor deze situatie. We vatten de 100 te trekken getallen op als realisaties van onafhankelijke stochasten $X_1, X_2, \ldots, X_{100}$, die uniform verdeeld zijn op $[0, 1]$, waarvan de verdelingsfunctie gegeven wordt door $\int_{-\infty}^{x} f(u)du$ met

$$f(u) = \begin{cases} 1 & , 0 \leq u \leq 1 \\ 0 & , \text{elders.} \end{cases}$$

Met andere woorden: voor elke X_i en $0 \leq x \leq 1$ geldt

$$\mathbf{P}(0 \leq X_i \leq x) = \int_0^x f(u)du = x.$$

Gevraagd wordt naar $\mathbf{P}(X_1 + X_2 + \cdots + X_{100} \geq 55)$. We geven een benadering voor deze kans met behulp van de centrale limietstelling.
De gevraagde kans is ook te formuleren als

$$\mathbf{P}(\bar{X}_{100} \geq 0.55).$$

Volgens de Centrale Limietstelling heeft

$$\sqrt{100}\frac{\bar{X}_{100} - m}{d}$$

bij benadering een standaard normale verdeling. Wanneer we m en d bepalen, kunnen we vervolgens de tabel van de standaard normale verdeling gebruiken om de kans uit te rekenen.
In dit geval zijn m en d de verwachting en standaarddeviatie van de uniforme verdeling op $[0, 1]$. Deze zijn gelijk aan :

$$m = \mathbf{E}(X_i) = \int_{-\infty}^{\infty} xf(x)dx = \int_0^1 xdx = \frac{1}{2}$$

en omdat

$$\mathbf{E}(X_i^2) = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^1 x^2 dx = \frac{1}{3},$$

is

$$d = \sqrt{\mathbf{E}(X_i^2) - \frac{1}{4}} = \frac{1}{\sqrt{12}}.$$

We hebben nu dat

$$\begin{aligned}
\mathbf{P}(\bar{X}_{100} \geq 0.55) &= \mathbf{P}\left(\sqrt{100}\frac{\bar{X}_{100} - 0.5}{1/\sqrt{12}} \geq \sqrt{100}\frac{0.55 - 0.5}{1/\sqrt{12}}\right) \\
&= \mathbf{P}\left(\sqrt{100}\frac{\bar{X}_{100} - 0.5}{1/\sqrt{12}} \geq 1.732\right) \\
&\approx \mathbf{P}(Z \geq 1.732),
\end{aligned}$$

waarbij Z standaard normaal verdeeld is. We vinden tenslotte in een tabel van de standaard normale verdeling dat $\mathbf{P}(Z \geq 1.732) = 0.0418$. Met andere woorden, de gevraagde kans is bij benadering gelijk aan 0.0418.

Normale benadering voor de binomiale verdeling.
Een stochast X, die binomiaal verdeeld is met parameters n en p kan geschreven worden als een som

$$X = X_1 + X_2 + \cdots + X_n$$

van n onafhankelijke stochasten met

$$\begin{aligned}
m &= \mathbf{E}(X_i) = p \\
d &= \sqrt{\mathrm{var}(X_i} = \sqrt{p(1-p)}.
\end{aligned}$$

Zij $\bar{X}_n = (X_1 + \cdots + X_n)/n$ het gemiddelde van X_1 tot en met X_n. Dan is $X = n\bar{X}_n$ en op grond van de Centrale Limietstelling is de stochast

$$\sqrt{n}\frac{\bar{X}_n - p}{\sqrt{p(1-p)}}$$

bij benadering standaard normaal verdeeld. Als we teller en noemer met $\sqrt{n}$ vermenigvuldigen vinden we dat

$$\frac{X - np}{\sqrt{np(1-p)}}$$

bij benadering standaard normaal verdeeld is.
Zij X binomiaal verdeeld met parameters $n = 100$ en $p = 0.25$. Men kan nagaan dat exact geldt

$$\mathbf{P}(X \leq 20) = 0.1488.$$

Als we de verdeling van X benaderen met de normale verdeling dan vinden we met $np = 25$ en $np(1-p) = 18.75$ dat:

$$\mathbf{P}(X \leq 20) = \mathbf{P}\left(\frac{X - 25}{\sqrt{18.75}} \leq \frac{20 - 25}{\sqrt{18.75}}\right) \approx \mathbf{P}(Z \leq -1.15).$$

waarbij Z standaard normaal verdeeld is, zodat

$$\mathbf{P}(Z \leq -1.15) = 0.1251.$$