Beraming van die kolfvermoë van Suid-Afrikaanse internasionale eendagkrieketspelers

H.H. Lemmer
Randse Afrikaanse Universiteit, Posbus 524, Aucklandpark, 2006
hhl@na.rau.ac.za

Ontvang 17 Januarie 2001; aanvaar 16 Februarie 2001

Uittreksel

Daar word aangetoon dat die internasionaal aanvaarde gebruik om 'n kolwer se vermoë te beraam deur sy gemiddeld gebaseer op sogenaamde voltooide beurte, dit wil sê sy totale getal lopies gedeel deur die aantal keer wat hy uit was, berus op die aanname dat sy tellings 'n eksponensiaalverdeling besit. Die meer algemene gamma- en Weibullverdelings gee egter beter passings en is ook meer sinvol om te gebruik aangesien nie-uit-nie-tellings op 'n meer bevredigende manier hanteer word. Beramings word gemaak van die S.A. internasionale eendagspelers gebaseer op data tot aan die einde van Augustus 2000. Daarvolgens is Klusener die beste kolwer met 'n beraamde gemiddeld heelwat hoër as Kirsten s'n, hoewel hulle volgens die gewone metode ongeveer dieselfde gemiddeld het. Die Weibullverdeling word aanbeveel as 'n baie geskikte model.

Abstract

Estimation of the batting ability of South African international one-day cricket players

It is shown that the international method of estimating a batsman's ability by the average number of runs per completed innings, i.e. the total number of runs divided by the number of out innings, follows from the assumption that the scores follow an exponential distribution. The more general gamma and Weibull distributions are better representatives of the scores and are more satisfactory because not out scores are treated in a more sensible way. Estimates are made for the South African international one-day players based on data up to the end of August 2000. It is found that Klusener is the best batsman with an estimated average markedly higher than that of Kirsten, although according to the usual method their averages are approximately the same. The Weibull distribution is recommended as a very suitable model.

1. INLEIDING

Laat $x_1, x_2, \ldots, x_n$ die tellings van 'n speler aandui waarby $x_1, x_2, \ldots, x_m$ sy uit-tellings is en $x_{m+1}, \ldots, x_n$ sy nie-uit-nie-tellings. Die teenwoordigheid van n.u.n. (nie uit nie)-tellings skep 'n probleem wanneer mens 'n statistiese verdeling by die data wil pas, omdat in elkeen van hierdie gevalle geredeneer kan word dat die speler waarskynlik 'n groter telling sou kon aanteken indien hy sy beurt kon voltooi. Die gebruiklike metode om verskillende verdelings by die data te pas en te toets watter een die beste is, kan hier nie gebruik word nie omdat sekere waarnemings gesensoreer is. Dit is dus nodig om van 'n maksimum aanneemlikheidsberaming, met inagneming van die gesensoreerde waarnemings, gebruik te maak.

2. BEPALING VAN DIE BESTE VERDELING

'n Ondersoek van al die tellings van die spelers onder beskouing het aangetoon dat die eksponensiaalverdeling sowel as die meetkundige verdeling redelike goeie passings lewer. Beskou eers die eksponensiaalverdeling met digtheidsfunksie

$$f_X(x) = \lambda e^{-\lambda x}, \quad x > 0.$$

Die aanneemlikheidsfunksie vir die datastel $x_1, \ldots, x_m, x_{m+1}, \ldots, x_n$ is

$$L = \prod_{i=1}^{m} \lambda e^{-\lambda x_i} \prod_{i=m+1}^{n} \left(\int_{x_i}^{\infty} \lambda e^{-\lambda u} du \right)$$

waarby vir n.u.n.-tellings die waarskynlikheid van 'n waarde $\geq x_i$ in berekening gebring word.

Deur op die gebruiklike wyse die maksimum aanneemlikheidsberamer te bereken, vind ons

$$\hat{\lambda} = \frac{m}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}_1}$$

waarby

$$\bar{x}_1 = \frac{1}{m} \sum_{i=1}^{n} x_i,$$

die gemiddeld gebaseer op voltooide beurte is.

Omdat die gammaverdeling 'n veralgemening van die eksponensiaalverdeling is, is dit vanselfsprekend dat dit 'n datastel beter sal beskryf. Beskou nou

$$f_X(x) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \cdot x^{\alpha-1} e^{-\lambda x} \quad \text{vir} \quad x > 0.$$

Indien beramers $\hat{\lambda}$ en $\hat{\alpha}$ deur middel van 'n passing van 'n gammaverdeling verkry word, kan die vermoë van 'n kolwer beraam word deur die gemiddeld van hierdie verdeling, naamlik $\hat{\alpha}/\hat{\lambda}$ te bereken. Die aanneemlikheidsfunksie vir 'n speler met tellings $x_1, \ldots, x_m, x_{m+1}, \ldots, x_n$ word gegee deur

$$L = \prod_{i=1}^{m} \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \cdot x_i^{\alpha-1} e^{-\lambda x_i} \prod_{i=m+1}^{n} \int_{x_i}^{\infty} \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \cdot u^{\alpha-1} e^{-\lambda u} du$$

$$= \frac{\lambda^{m\alpha}}{[\Gamma(\alpha)]^m} \prod_{i=1}^{m} x_i^{\alpha-1} \cdot e^{-\lambda \sum_{i=1}^{m} x_i} \prod_{i=m+1}^{n} \left[1 - \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \int_0^{x_i} u^{\alpha-1} e^{-\lambda u} du \right].$$

Die bepaling van die maksimum aanneemlikheidsberamers is nou geen maklike taak nie omdat die afgeleides van *L* (sowel as van *L* self) funksies bevat waarvoor eksplisiete uitdrukkings nie gevind kan word nie. Vir elke speler is die beramings $\hat{\alpha}$ en $\hat{\lambda}$ gevind deur die funksie *L* rekenaarmatig met betrekking tot α en λ gesamentlik te maksimeer. *L* is eerstens oor 'n rooster van (x, λ)-waardes bereken. Daarna is die rooster telkens oor die optimale gebied verfyn totdat die akkuraatheid van die parameters tot vier desimale syfers verkry is. Kyk tabel 1 vir die parameterwaardes.

Die beraamde gemiddeldes verkry volgens die gammamodel, naamlik $\bar{x}_b$,word in tabel 2 gegee. Daarin verskyn ook die gemiddeld $\bar{x}_1$, die gemiddeld $\bar{x}_u$ gebaseer slegs op die uit-tellings en die gemiddeld $\bar{x}^*$ gebaseer slegs op n.u.n.-tellings. Dit is opmerklik dat in die geval van al die goeie kolwers $\bar{x}^*$ baie hoër is as $\bar{x}_u$. Hierdie is veral 'n funksie van die kolfposisie van die kolwer. 'n Mens sou verwag dat die kolwer wat vroeg gaan kolf minder n.u.n.-tellings sal behaal en dan ook 'n hoër gemiddelde n.u.n.-telling sal hê. 'n Mens kan tereg redeneer dat as 'n beurt eindig en die kolwer nie uit is nie, hy na verwagting nog meer lopies sou kon aanteken indien hy sou kon voortkolf. Dit is dus sinvol om in die geval van n.u.n.-tellings 'n beraming te maak van watter telling die kolwer na verwagting sou kon kry, indien hy sou voortkolf totdat hy uit is. Omdat ons die kolwer se verdeling bepaal het, kan ons sy n.u.n.-tellings aanvul tot sy verwagte tellings, gegee dat hy alreeds x_i^* lopies aangeteken het. Ons bereken

$$E\left(X|X \geq x_i^*\right) = \frac{1}{P\left(X \geq x_i^*\right)} \int_{x_i^*}^{\infty} x f_X\left(x\right) dx$$

waar X 'n gamma (x, λ) -verdeling besit met x en λ bekend.

Nou is

$$P\left(X \geq x_i^*\right) = 1 - \int_0^{x_i^*} f_X\left(x\right) dx$$

met die integraal 'n onvolledige gammafunksie. Daar kan aangetoon word dat

$$\int_{x_i^*}^{\infty} x f_X\left(x\right) dx = \frac{\alpha}{\lambda}\left(1 - \int_0^{x_i^*} f_Y\left(y\right) dy\right)$$

met Y 'n gamma ($\alpha + 1$, λ)-variaat.

Vir elke speler is elke n.u.n.-telling vervang deur die beraamde telling. Hierdie is by die uit-tellings getel en die totaal is gedeel deur n. Sodoende is 'n verdere beraming $\bar{x}_a$ gekry van die speler se vermoë – kyk tabel 2. Vergelyk ons $\bar{x}_b$ en $\bar{x}_a$, is dit duidelik dat daar 'n baie groot ooreenstemming is. Dit bevestig dat die gepaste modelle en beramings van die spelers se vermoëns inderdaad baie goed is en dat die "aanvulling" van n.u.n.-tellings sinvol is.

Die Weibullverdeling is ook 'n veralgemening van die eksponensiaalverdeling en het vir geskikte parameterwaardes 'n vorm wat passend is vir kriekettellings.

Beskou

$$f\left(x\right) = \beta\lambda^{\beta} x^{\beta-1} e^{-(\lambda x)^{\beta}}, \;\; x > 0, \; \beta > 0, \; \lambda > 0.$$

Die aanneemlikheidsfunksie is

$$L = \prod_{i=1}^{m} \beta\lambda^{\beta} x_i^{\beta-1} eksp\left(-\left(\lambda x_i\right)^{\beta}\right) \prod_{i=m+1}^{n} \int_{x_i}^{\infty} \beta\lambda^{\beta} u^{\beta-1} eksp\left(-\left(\lambda u\right)^{\beta}\right) du$$

wat geskryf kan word as

$$L = \beta^m \lambda^{m\beta} \prod_{i=1}^{m} x_i^{\beta-1} eksp\left[-\sum_{i=1}^{n}\left(\lambda x_i\right)^{\beta}\right].$$

Die optimale keuse van β en λ kan ook rekenaarmatig gevind word en daaruit die beraamde gemiddeld vir die kolwer

$$\bar{x}_b = \left(\hat{\lambda}\right)^{-1/\hat{\beta}} \cdot \Gamma\left(1 + 1/\hat{\beta}\right).$$

Aangevulde tellings word beraam deur gebruik te maak van

$$E\left(X|X \geq x_i\right) = \frac{1}{P\left(X \geq x_i\right)} \int_{x_i}^{\infty} \beta\lambda^{\beta} x^{\beta} eksp\left(-\left(\lambda x\right)^{\beta}\right) dx$$

waar

$$P\left(X \geq x_i\right) = eksp\left(-\left(\lambda x_i\right)^{\beta}\right)$$

en

$$\int_{x_i}^{\infty} \beta\lambda^{\beta} x^{\beta} eksp\left(-\left(\lambda x\right)^{\beta}\right) dx = \frac{1}{\lambda} \int_{(\lambda x_i)^{\beta}}^{\infty} u^{\frac{1}{\beta}} e^{-u} du.$$

Hieruit kan die aangevulde tellings rekenaarmatig bereken word en die aangevulde gemiddeld $\bar{x}_a$ gevind word. Die resultate verskyn ook in tabel 2 en die parameterwaardes in tabel 1.

3. INTERPRETASIE VAN RESULTATE

Uit teoretiese en praktiese oorwegings is die gammaverdeling en die Weibullverdeling beter modelle as die eksponensiaalverdeling. By die meeste kolwers lewer die gamma- en Weibullverdeling beraamde gemiddeldes wat baie dieselfde is. Die ooglopende uitsondering is Klusener, wat met die Weibullverdeling 'n heelwat hoër gemiddeld het as met die gammaverdeling. Ten einde te kan kies tussen hierdie twee verdelings, is 'n gammaverdeling gepas op die gewone en aangevulde tellings wat vanuit 'n gammapassing bereken is, en dieselfde is gedoen met die Weibullverdeling. Deur toepassing van die Kolmogorov-Smirnov-passingstoets[1] blyk dit dat in die geval van al die spelers die Weibullverdeling 'n beter passing gee as die gammaverdeling – kyk tabel 3. Dit is dus duidelik dat die gemiddeldes $\bar{x}_b$ (Weibull) die beste aanduiding gee van die kolwers se vermoëns.

Die Weibullverdeling is 'n leeftydverdeling waarvan die oombliklike falingskoers *(instantaneous failure rate)* gegee word deur[2]

$$h\left(x\right) = f\left(x|X \geq x\right) = \beta\lambda\left(\lambda x\right)^{\beta-1} \;\; ; \;\; x > 0, \, \beta > 0$$

Vir $\beta < 1$ is die falingskoers dalend. Dit beteken dat hoe hoër 'n kolwer se telling is, hoe kleiner is die waarskynlikheid dat hy sy paaltjie sal verloor. In die geval van Klusener is $\beta = 0.6795$ en weet almal dat as hy eers op dreef is, hy moeilik uitgekry word. Uit 'n ondersoek van 'n groot aantal internasionale eendagspelers blyk dit dat in 'n baie hoë persentasie van tellings bo 100, die speler nie uit was nie. Dit ondersteun die bevinding dat die Weibullverdeling 'n goeie model is vir kriekettellings van eendagkolwers.

4. SLOTOPMERKINGS

4.1 Omdat kriekettellings diskreet is, kan gevra word of daar nie liewer met diskrete verdelings gewerk moet word nie. Die meetkundige verdeling (wat as die diskrete ekwivalent van die eksponensiaalverdeling beskou kan word), lewer presies dieselfde beramer vir die gemiddeld, naamlik $\overline{x}_1$. Die logiese veralgemening van die meetkundige verdeling, naamlik die negatiewe binomiaalverdeling, is ook op soortgelyke wyse as tevore op die data gepas, maar sy optimale keuse van parameterwaardes het in elke geval herlei na die meetkundige verdeling (omdat die negatiewe binomiaalverdeling se tweede parameter diskreet is). Dit was dus nie 'n vrugbare oefening nie.

4.2 Uiteraard speel baie faktore 'n rol by die prestasie van 'n kolwer en is dit onmoontlik om alles in ag te neem by die beoordeling van sy vermoë. Aanvangskolwers moet die aanslag van uitgeruste boulers wat 'n nuwe bal gebruik, trotseer, maar hulle het meer tyd om 'n lang beurt te speel as 'n kolwer wat op nommer 5 inkom. Die gemiddeldes wat beraam is, moet dus teen hierdie agtergrond beoordeel word. 'n Alternatiewe maatstaf wat gebruik kan word, is die getal lopies aangeteken per 100 balle ontvang (die sogenaamde "*strike rate*"). Hierop word nie in hierdie studie ingegaan nie.

4.3 'n Alternatiewe metode wat gebruik kan word, is om van leeftydverdelings gebruik te maak. Dit kom neer op die passing van 'n leeftydverdeling op die gesensoreerde data deur byvoorbeeld van die Kaplan-Meier-beramer gebruik te maak. Hierdie metode stel mens in staat om die waarskynlikheid te beraam dat 'n kolwer meer as 'n spesifieke getal lopies sal aanteken. Daar word egter nie verder hierop ingaan nie.[3]

4.4 Die finale gevolgtrekking is dat die tradisionele gemiddeld (gebaseer op 'n eksponensiaalverdeling as model), wat maklik berekenbaar is, 'n redelike aanduiding van die meeste kolwers se vermoëns gee, maar dat die Weibull-gemiddeld 'n meer betroubare maatstaf is.

Tabel 1 Parameterwaardes vir aangepaste verdelings

	Gammaverdeling		Weibullverdeling	
Speler	$\hat{\alpha}$	$\hat{\lambda}$	$\hat{\beta}$	$\hat{\lambda}$
Cullinan	0.9433	0.0276	0.9855	0.0295
Klusener	0.6795	0.0144	0.7426	0.0235
Kallis	0.8701	0.0211	0.9206	0.0250
Kirsten	0.7654	0.0182	0.8415	0.0256
Rhodes	1.041	0.0338	1.0447	0.0321
Pollock	0.8314	0.0318	0.8837	0.0401
Boucher	0.6957	0.0374	0.7511	0.0623

Tabel 3 P-waardes van Kolmogorov-Smirnovtoets

Speler	Gammaverdeling	Weibullverdeling
Cullinan	0.565	0.587
Klusener	0.161	0.234
Kallis	0.439	0.644
Kirsten	0.263	0.310
Rhodes	0.158	0.176
Pollock	0.012	0.041
Boucher	0.720	0.768

Tabel 2 Prestasies van Suid-Afrikaanse spelers tot einde Augustus 2000

Speler	**Gem. uit** $\overline{x}_u$ **(aantal)**	**Gem. n.u.n.** $\overline{x}^*$ **(aantal)**	$\overline{x}_1$	**Gamma-verdeling**		**Weibull-verdeling**	
				x_b	x_a	x_b	x_a
Cullinan	27.89 (113)	43 (16)	33.97	34.14	34.14	34.11	34.07
Klusener	21.18 (56)	41.04 (28)	41.70	47.19	47.17	51.11	50.98
Kallis	30.04 (86)	53.18 (17)	40.55	41.24	41.25	41.50	41.44
Kirsten	33.10 (124)	71.86 (14)	41.21	42.06	42.07	42.71	42.52
Rhodes	22.87 (131)	33.16 (32)	30.97	30.80	30.81	30.60	30.62
Pollock	16.70 (54)	17.15 (26)	24.96	26.14	26.12	26.5	24.45
Boucher	13.17 (41)	13.92 (13)	17.59	18.60	18.60	19.09	19.26

LITERATUURVERWYSINGS

1. Conover, W.J. (1980). *Practical Nonparametric Statistics*. 2nd ed. (J. Wiley & Sons Inc.) pp. 345-349.
2. Bain, L.J., Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistics*. 2nd ed. (PWS-KENT Publishing Company) p. 542.
3. Elandt-Johnson, R.C., Johnson, N.L. (1980). Survival Models and Data Analysis. (J. Wiley & Sons Inc.) pp. 172-173.

HOFFIE LEMMER

Hoffie Lemmer is professor in Statistiek aan die Randse Afrikaanse Universiteit, waar hy vir meer as dertig jaar werksaam is. Hy het aan die Universiteit van Pretoria gestudeer en het die graad D.Sc. aldaar verwerf. Hy het ook daar sy beroepsloopbaan begin as lektor en later senior lektor, voordat hy by die R.A.U. as professor aangestel is. Hy is lid van die S.A. Statistiese Vereniging, die S.A. Akademie vir Wetenskap en Kuns en die Institute of Mathematical Statistics. Hy is ’n geregistreerde natuurwetenskaplike. Sy hoof navorsingsrigting is nieparametriese statistiek, maar sy navorsingsaktiwiteite strek wyer – selfs tot by sportstatistiek as gevolg van ’n besondere belangstelling in krieket.