

Odds en waarskynlikhede met toepassings

Pieta van Deventer

Trefwoorde: aanneemlikheid, aanneemlikheidsverhouding, deltametode, kansverhouding, plasebo, relatiewe kansverhouding, vertrouensinterval, voorwaardelik, waarskynlikheid

Weet ons werklik wat die woord *odds*, wat baie algemeen, veral in gesprekke oor perdewedrenne, gebruik word, beteken?

Die begrippe *odds* en *odds ratio* versus *waarskynlikheid* word bespreek deur eers twee voorbeelde te gebruik en dit dan in algemene terme uit te druk.

'n Statisties verantwoordbare vertrouensinterval vir betekenisvolle waardes word gegee en aan die einde word die ietwat lastige afleiding van laasgenoemde in 'n ekstra afdeling voorsien. Na dese sal jy meer as die meeste mense weet van odds en wat dit regtig beteken. Baie mense gebruik die woord gereeld, maar weet hulle regtig waarom dit gaan? Dis te betwyfel.

Ons hoor dikwels iets soos die volgende: “Die odds vir 'n perd om te wen is 1 tot 11”, of “Die odds is 1 teen 4 dat dit voor sesuur vanaand gaan reën.” Die bedoeling is dan dat, gegee soortgelyke omstandighede as wat in die verlede gegeld het, die waarskynlikheid 1/11 is dat die perd sal wen en 1/4 is dat dit voor vanaand 18:00 sal reën.

Die woord *odds* word dus ongelukkig dikwels in die foutiewe sin van “waarskynlikheid” soos hier bo gebruik. Daar is wél waarskynlikhede betrokke as jy dit sou wou gebruik, maar dit is nie 1/11 of 1/4 nie. Gegewe bostaande odds is die betrokke waarskynlikhede wel 1/12 en 0,2. Dit is duidelik dat die odds uit 'n kombinasie van die teller en noemer van die betrokke waarskynlikheid bestaan. In hierdie geval is die waarskynlikhede $1/(1 + 11) = 1/12$ en $1/(1 + 4) = 1/5$.

In hierdie bydrae word kortliks gekyk na die werklike betekenis van *odds*, wat die ooreenstemmende Afrikaanse benaming daarvoor is en hoe dit toegepas word. Kortliks is die odds op die ou einde niks anders as die verhouding tussen twee getalle nie, maar dis nie 'n waarskynlikheid nie. Die woord *odds* word dan ook vertaal met die woord *aanneemlikheid*. Later sal ons sien dat daar wel op 'n manier 'n waarskynlikheidsgedagte aan gekoppel kan word wat dan tot 'n nuwe benaming lei.

Verder word die term *odds ratio* ook gereeld gebruik. Hiervoor gebruik ons die term *aanneemlikheidsverhouding*. Dis hiér waar die verwarring dikwels ontstaan, selfs by ervare statistici wat nie elke dag daarmee werk nie, naamlik tussen *aanneemlikheid* (odds) wat 'n verhouding verteenwoordig en die *aanneemlikheidsverhouding* (odds ratio) wat dus die verhouding tussen twee verhoudings is.

Grieks? Nee, nie regtig nie. Deur die volgende voorbeeld te bestudeer, sal jy sien dat dit nie net redelik duidelik is nie, maar ook baie nuttig, want dit plaas bepaalde verhoudings in perspektief. Die voorbeeld wat gebruik word, kom voor in Alan Agresti se boek¹ wat saam

¹ Agresti, Alan (2002), *Categorical data analysis*, 2de uitgawe, Wiley-Interscience, A Wiley & Sons, Inc, Hoboken, New Jersey.

met die boek van Bishop, Fienberg en Holland² sekerlik van die netjiesste en beste beskrywings van hierdie begrippe gee. Daar is vele naslaanboeke oor hierdie en verwante onderwerpe, maar hierdie twee boeke bly nog die beste waarde vir geld, alhoewel Agresti se boek jou 'n paar duisend rand uit die sak sal jaag. Die data wat Agresti gebruik, kom uit "Preliminary report: Findings from the aspirin component of the ongoing physicians' health study, *New England Journal of Medicine*, 318:262–4 (1988).

Kruisklassifikasie van aspiriengebruik versus miokardiale infarksie			
	Miokardiale infarksie		
	Noodlottige aanval	Nienoodlottige aanval	Geen aanval nie
Plasebo	18	171	10 845
Aspirien	5	99	10 933

Hierdie data is uit 'n studie onder 21 778 medici verkry. Die doel van die studie was om die hipotese te toets dat die daaglikse toediening van aspirien die voorkoms van hartaanvalle in 'n bevolking in die algemeen behoort te verminder.

Aanvanklik beskou ons slegs 'n verwerking van bostaande tabel, naamlik:

Kruisklassifikasie van aspiriengebruik versus miokardiale infarksie		
	Miokardiale infarksie	
	Hartaanval – ongeag graad van aanval	Geen aanval nie
Plasebo	189	10 845
Aspirien	104	10 933

Nota: Die notasie wat hier gebruik word, is ongelukkig nie universeel nie. Waar θ hier gebruik word, word in die literatuur soms die Ω -simbool gebruik en omgekeerd. Die voetskrif *ij* hieronder verwys na ry *i* en kolom *j*.

Die aanneemlikheid³ (odds) van 'n hartaanval versus geen hartaanval onder plasebogebruikers is 189:10845, wat as $\Omega_{11} = \frac{189}{10845} = 0.01743$ geskryf word. Let op dat hierdie waarde nie 'n waarskynlikheid is nie.

Let op die volgende omgekeerde argument: Die aanneemlikheid (odds) van geen hartaanval versus 'n hartaanval onder plasebogebruikers is 10845:189, wat as $\Omega_{12} = \frac{10845}{189} = 57.381$ geskryf word. Dis gevolglik duidelik dat ons hier geen waarskynlikheid het nie.

Voer nou die aanvanklike bewering verder. Indien hierdie steekproef as verteenwoordigend van die populasie beskou kan word, is die waarskynlikheid dat 'n persoon 'n aanval sal kry as

² Bishop, YMM, SE Fienberg en PW Holland (1975), *Discrete multivariate analysis*, Cambridge, MA, MIT Press.

³ Hierdie verwysing na aanneemlikheid moet nie met die begrip aanneemlikheid (*likelihood*) soos toegepas deur Fisher en Neymann, twee van die vaders van wiskundige statistiek soos ons dit vandag ken, verwar word nie. Vir meer oor hierdie heftige stryd, want dit was 'n stryd tussen reuse, verwys gerus na <https://www.jstor.org/stable/3541653>

hy/sy die plasebo gebruik, benaderd $p_{11} = \frac{189}{189+10845} = \frac{189}{11034} = 0.01713$. Hierteenoor, weer onder die aanname van 'n verteenwoordigende steekproef, is die aanneemlikheid vir 'n aspiriengebruiker om 'n hartaanval te kry 104:10933, dws $\Omega_{21} = \frac{104}{10933} = 0.009511$ en die waarskynlikheid dat 'n persoon 'n aanval sal kry as hy/sy 'n aspiriengebruiker is, benaderd $p_{21} = \frac{104}{104+10933} = \frac{104}{11037} = 0.009423$.

Ons wil nou Ω_{11} vergelyk met Ω_{21} . Hiervoor word die verhouding van die verhoudings gebruik, naamlik die aanneemlikheidsverhouding (odds ratio), θ . In hierdie geval is

$$\theta = \frac{\Omega_{11}}{\Omega_{21}} = \frac{189/\cancel{10845}}{104/\cancel{10933}} = \frac{189 \cdot 10933}{104 \cdot 10845} = \frac{0.01743}{0.00951} = 1.83281.$$

Dit beteken dat die plasebogebruiker se aanneemlikheid om 'n hartaanval te kry 83% hoër/meer is as die aanneemlikheid van die aspiriengebruiker om 'n hartaanval te kry. Anders gestel, die plasebogebruiker se aanneemlikheid om 'n hartaanval te kry, is 1,83 keer dié van die aspiriengebruiker (nie 1,83 *meer* nie, maar 1,83 *maal* ..., of 1,83 *soveel as* ...).

Die aanneemlikheidsverhouding (odds ratio) stel ons dus in staat om sin te maak van die twee individuele aanneemlikhede deur hulle te vergelyk dmv 'n verhouding.

Beskou die oorspronklike data nou in meer besonderhede deur na 'n subtabel te kyk om verdere afleidings te kan maak:

Indien 'n persoon 'n hartaanval kry, kan ons nou verder sê dat die aanneemlikheid van 'n noodlottige hartaanval versus 'n nienoodlottige hartaanval onder plasebogebruikers 18:171 is, wat as $\Omega_{11} = \frac{18}{171}$ geskryf word. Let weer eens op dat dit nie 'n waarskynlikheid is nie. Indien

hierdie steekproef as verteenwoordigend van die populasie beskou word, is die waarskynlikheid dat 'n persoon wat 'n aanval kry se aanval noodlottig sal wees as hy/sy die plasebo gebruik het, benaderd $p_{11} = \frac{18}{189}$. Hierteenoor, weer eens onder die aanname van 'n verteenwoordigende steekproef, is die aanneemlikheid vir 'n aspiriengebruiker wat 'n aanval kry dat dit noodlottig sal wees, 5:99, dws $\Omega_{21} = \frac{5}{99}$ en die waarskynlikheid dat 'n persoon wat 'n aanval kry as hy/sy 'n aspiriengebruiker is dat die aanval noodlottig sal wees, benaderd $\frac{5}{104}$.

Die aanneemlikheidsverhouding dat 'n aanval, gegewe 'n hartaanval, noodlottig sal wees by

'n plasebogebruiker versus 'n aspiriengebruiker is $\theta = \frac{\Omega_{11}}{\Omega_{21}} = \frac{18/\cancel{171}}{5/\cancel{99}} = \frac{18 \cdot 99}{5 \cdot 171} = 2.084$, dws

ongeveer twee keer *soveel as* vir die aspiriengebruiker, let wel, nie twee keer *meer* soos sommige dit bewoord nie. Die bewoording is: ... *twee keer soveel as*

Die skyn asof daar werklik 'n voordeel in is om maar liewer die aspirien te gebruik.

Dis duidelik dat indien die aanneemlikheidsverhouding gelyk aan 1 is, daar geen voordeel in is om die aspirien te gebruik nie. Hoe weet ons of die syfers 1,83281 en 2,084 statisties betekenisvol van 1 afwyk? Wie sê dis nie maar net toevallig nie? Hiervoor benodig ons 'n vertrouensinterval vir θ .

Let op $0 \leq \theta \leq \infty$ met intervallengtes $[0,1] \ll [1,\infty]$ weerskante van die gelykbreekpunt van 1, wat 'n hewige skewe verdeling van waardes met negatiewe waardes teenoor positiewe waardes impliseer ($\ll$ beteken “baie kleiner as”). 'n Waarde groter as 1 beteken dat die beslissing na die teller se herkoms gaan. Omgekeerd as die waarde kleiner is as 1 gaan die beslissing na die noemer se herkoms. Hierdie skewe verdeling kan gelukkig bereedder word deur 'n logaritmiese transformasie van θ . Dit kan mbv die deltametode wat by 'n vorige geleentheid onder 'n bespreking van toepassings van Taylor se reeks⁴ bespreek is, aangetoon word dat 'n logaritmiese transformasie van $\hat{\theta}$ benaderd soos volg verdeel is:

$$\ln \hat{\theta} \sim N\left(\ln \theta, \sum_i \sum_j \frac{1}{n_{ij}}\right)$$

Die bewys van hierdie resultaat is ietwat lastig, maar as jy van 'n uitdaging hou, nooi ek jou uit om die afdeling “Vir verryking” onderaan hierdie artikel te bestudeer. As jy daarvan hou, wil ek jou geluk wens, want dan is jy reeds 'n kandidaat vir 'n studie in wiskundige statistiek. Die krag van logiese denke en wiskundige elegansie is heerlik om mee te heul. Sterkte.

Pas nou hierdie resultaat toe op ons voorbeelde hier bo: Let daarop dat al ons resultate eintlik ramings is en dus met 'n kappie voorgestel moes word. Vir die doel van eenvoud is dit tot nou toe weggelaat, maar van hier af is dit wel van belang.

Dit beteken dat ons $(1-\alpha)100\%$ vertrouensintervalle vir θ as volg kan skep, sê maar vir 'n aanvaarde vertroue van 95% dat ons interval die werklike onbekende waarde van θ sal bevat, of anders gestel, vir $\alpha = 0.05$:

$$\ln \hat{\theta} \pm z_{0.05/2} \left(\sum_i \sum_j \frac{1}{n_{ij}} \right)^{1/2}$$

Pas hierdie berekeninge nou op die voorbeelde hier bo toe, dan vinde ons dat die 95% vertrouensintervalle vir θ soos volg lyk:

$$L_{bo} = \exp \left(\ln \hat{\theta} + z_{0.025} \left(\sum_i \sum_j \frac{1}{n_{ij}} \right)^{1/2} \right)$$

$$L_{onder} = \exp \left(\ln \hat{\theta} - z_{0.025} \left(\sum_i \sum_j \frac{1}{n_{ij}} \right)^{1/2} \right)$$

⁴ <https://www.litnet.co.za/wiskunde-met-pieta-van-deventer-taylor-se-reeks>

Voorbeeld 1

Kruisklassifikasie van aspiriengebruik versus miokardiale infarsie		
	Miokardiale infarsie	
	Hartaanval ongeag graad	Geen aanval nie
Plasebo	189	10 845
Aspirien	104	10 933

$$\begin{aligned}\text{Limiet}_{bo} &= \exp \left(\ln \hat{\theta} + z_{\alpha/2} \left(\sum_i \sum_j \frac{1}{n_{ij}} \right)^{1/2} \right) \\ &= \exp \left(\ln 1.83281 + 1.96 \cdot \left(\frac{1}{189} + \frac{1}{10845} + \frac{1}{104} + \frac{1}{10933} \right)^{1/2} \right) \\ &= 2.3308\end{aligned}$$

$$\begin{aligned}\text{Limiet}_{onder} &= \exp \left(\ln \hat{\theta} - z_{\alpha/2} \left(\sum_i \sum_j \frac{1}{n_{ij}} \right)^{1/2} \right) \\ &= \exp \left(\ln 1.83281 - 1.96 \cdot \left(\frac{1}{189} + \frac{1}{10845} + \frac{1}{104} + \frac{1}{10933} \right)^{1/2} \right) \\ &= 1.44\end{aligned}$$

Hierdie interval sluit nie die waarde 1 in nie. Die gevolgtrekking is dat die waarde van θ waarskynlik nie 1 is nie en dat die gebruik van aspirien dus betekenisvol is om 'n verskil te maak tov die aanneemlikheid om 'n potensiële hartaanval af te weer.

Voorbeeld 2

Kruisklassifikasie van aspiriengebruik versus miokardiale infarsie		
	Miokardiale infarsie	
	Noodlottige aanval	Nienoodlottige aanval
Plasebo	18	171
Aspirien	5	99

$$\begin{aligned}L_{bo} &= \exp \left(\ln \hat{\theta} + z_{\alpha/2} \left(\sum_i \sum_j \frac{1}{n_{ij}} \right)^{1/2} \right) \\ &= \exp \left(\ln 2.084 + 1.96 \cdot \left(\frac{1}{18} + \frac{1}{171} + \frac{1}{5} + \frac{1}{99} \right)^{1/2} \right) \\ &= 5.787\end{aligned}$$

$$\begin{aligned}
L_{\text{onder}} &= \exp \left(\ln \hat{\theta} - z_{\alpha/2} \left(\sum_i \sum_j \frac{1}{n_{ij}} \right)^{1/2} \right) \\
&= \exp \left(\ln 2.084 - 1.96 \cdot \left(\frac{1}{18} + \frac{1}{171} + \frac{1}{5} + \frac{1}{99} \right)^{1/2} \right) \\
&= 0.7506
\end{aligned}$$

Hierdie interval sluit die waarde 1 in. Die gevolgtrekking is dat die gebruik van aspirien oënskynlik nie 'n beduidende verskil in die graad van 'n hartaanval, as jy 'n aanval sou kry, sal maak nie.

Hoe sou ons bostaande kan veralgemeen?

Indien ons die tabelle vir 'n eenvoudige geval sou veralgemeen, sien dit soos volg daar uit:

	Uitkoms 1	Uitkoms 2	Totaal
Scenario 1	n_{11} $p_{11} = \frac{n_{11}}{n_{1+}}$	n_{12} $p_{12} = \frac{n_{12}}{n_{1+}}$	$n_{11} + n_{12} = n_{1+}$
Scenario 2	n_{21} $p_{21} = \frac{n_{21}}{n_{2+}}$	n_{22} $p_{22} = \frac{n_{22}}{n_{2+}}$	$n_{21} + n_{22} = n_{2+}$

Hier dui n_{ij} 'n waargenome frekwensie met uitkoms j , gegewe scenario i aan en so ook dui p_{ij} 'n waargenome waarskynlikheid met uitkoms j , gegewe scenario i aan.

Dis volgende uitdrukkings toon die verwantskap tussen aanneemlikhede, die aanneemlikheidsverhouding en waarskynlikhede aan.

$$\begin{aligned}
\Omega_1 &= \frac{n_{11}}{n_{12}} = \frac{n_{11}/n_{1+}}{n_{12}/n_{1+}} = \frac{p_{11}}{p_{12}} \\
\Omega_2 &= \frac{n_{21}}{n_{22}} = \frac{n_{21}/n_{2+}}{n_{22}/n_{2+}} = \frac{p_{21}}{p_{22}} \\
\theta &= \frac{\Omega_1}{\Omega_2} = \frac{p_{11}/p_{12}}{p_{21}/p_{22}} = \frac{p_{11}p_{22}}{p_{21}p_{12}}
\end{aligned}$$

Dis die rede waarom die aanneemlikhede en die aanneemlikheidsverhoudings ook as die *kansverhoudings* en *relatiewe kansverhoudings* bekend staan. Dit is nog 'n rede waarom sommige dink dat *odds/aanneemlikhede* gelyk staan aan *waarskynlikhede*. Hierdie formaat van die tabel stel ons in staat om 'n vertrouensinterval vir θ af te lei soos in die afdeling “Vir verryking” hier onder uiteengesit.

Verder sou 'n mens ook die aanneemlikhede soos volg kon skryf, afhangende van of 'n mens uitkoms 1 of uitkoms 2 wil uitlig as moontlike “voorkeur”-gebeurtenis:

$\Omega_{11} = \frac{1}{\Omega_{12}}$ en so ook $\Omega_{21} = \frac{1}{\Omega_{22}}$. Dit sou dan beteken dat die aanneemlikheidsverhouding self

ook in inverse vorm kan voorkom. Dit beteken dus dat 'n mens baie versigtig met jou keuse van teller moet wees, gedagtig aan in watter uitkoms jy in die besonder belangstel.

As 'n nuttige nagedagte, net 'n vriendelike waarskuwing: In die uiteensettings hier bo is die ry-gegewens as die voorwaardelike velde hanteer en die kolom-gegewens as die resultate gegewe die ry-voorwaardes. Ongelukkig vind 'n mens die omgekeerde situasie in sommige boeke, veral in die veld van sielkunde, waar 'n mens dan ipv om van links na bo te redeneer jy van bo na links onder moet redeneer. Ja, dit maak dit lastig. Dis egter nie verkeerd nie. Dit beteken net dat 'n mens oopkop na die voorstellingswyse van die dokument waarmee jy besig is moet kyk, anders kan dit baie verwarrend raak.

Vir verryking

Hier kry ons te doen met 'n toepassing van die deltametode op 'n vektorfunksie van 'n asimptoties-normaal-ewekansige vektor. Die stelling is soos volg:

Beskou vektorfunksies van 'n asimptoties-normaal-ewekansige vektor met N parameters:

Laat $\underline{\theta} = (\theta_1, \theta_2, \dots, \theta_N)'$.

Laat $\underline{t} = \underline{\theta}$, en bereken $\underline{g}(\underline{t}) = (g_1(\underline{t}), g_2(\underline{t}), \dots, g_q(\underline{t}))'$, dan volg dit dat (let op dat q op die aantal funksies oftewel elemente van die vektor van funksies dui)

$$\left(\frac{\partial \underline{g}}{\partial \underline{t}} \right) \bigg|_{\underline{t}=\underline{\theta}} \bigg|_{q \times N} = \left(\frac{\partial g_i(\underline{t})}{\partial t_j} \right) \bigg|_{t_j=\theta_j} \bigg|_{q \times N}$$

$$\Rightarrow \sqrt{n}(\underline{g}(\underline{T}_n) - \underline{g}(\underline{\theta})) \sim N \left(\underline{0}, \left(\frac{\partial \underline{g}}{\partial \underline{\theta}} \right) \Sigma \left(\frac{\partial \underline{g}}{\partial \underline{\theta}} \right)' \right)$$

Die aantal rye en kolomme lyk soos volg: $\Theta_{q \times N} \quad \Sigma_{N \times N} \quad \Theta_{N \times q}$

Dit dui dus op een of meer funksies wat waargeneem word en dan na 'n vektor getransformeer word. Die n in $\underline{T}_n$ dui op 'n steekproef van grootte n .

Ons beskou nou hier die asimptoties-normaalverdeling van die aanneemlikheidsverhouding, ook bekend as die relatiewe kansverhouding. Dis bekend dat die konvergensie na normaliteit op die logaritmiese skaal vinniger is as op die basisverdeling, in die besonder waar ons hier met 'n uiters skewe verdeling te doen het waar die sentrale waarde 1 is, maar die definisiegebied van 0 tot ∞ strek.

Daar is drie stappe van belang, naamlik: Eerstens begin ons met 'n vektor van die selwaardes. Dit word later na 'n nuwe vektor getransformeer. Tweedens volg die logaritmiese transformasie en derdens 'n verskuiwing (*shift*) dmv 'n multiplikatiewe faktor om sodoende

die logkansverhoudings te bewerkstellig. Let op dat die volgende notasie vir die waargenome waardes gebruik word:

$$1. \quad \underline{p} = (p_{11} = p_1; p_{12} = p_2, p_{21} = p_3, p_{22} = p_4)$$

en dieselfde geld vir die onbekende selwaarskynlikhede, naamlik:

$$\underline{\pi} = (\pi_{11} = \pi_1; \pi_{12} = \pi_2, \pi_{21} = \pi_3, \pi_{22} = \pi_4)$$

Die kappie (^) hier onder dui soos gewoonlik op 'n beraming van 'n populasieparameter gebaseer op die waargenome steekproefwaardes.

2. Die algemene logtransformasievektor $\underline{g}(\cdot)$ van die vektor van die onbekende selwaarskynlikhede $\underline{\pi}$, naamlik $\underline{g}(\underline{\pi}) = \log \underline{\pi}$ en sy afgeleide word soos volg geskryf (N dui op 'n potensiële aantal selle wat in ons geval sleg 4 is en $q = 1$, aangesien ons slegs een funksie het.:

$$\begin{aligned} \frac{\partial \underline{g}}{\partial \underline{\pi}} &= \frac{\partial}{\partial \underline{\pi}} \begin{pmatrix} \log \pi_1 \\ \log \pi_2 \\ \vdots \\ \log \pi_N \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial \log \pi_1}{\partial \pi_1} & \frac{\partial \log \pi_1}{\partial \pi_2} & \dots & \frac{\partial \log \pi_1}{\partial \pi_N} \\ \frac{\partial \log \pi_2}{\partial \pi_1} & \frac{\partial \log \pi_2}{\partial \pi_2} & \dots & \frac{\partial \log \pi_2}{\partial \pi_N} \\ \dots & \dots & \dots & \dots \\ \frac{\partial \log \pi_N}{\partial \pi_1} & \frac{\partial \log \pi_N}{\partial \pi_2} & \dots & \frac{\partial \log \pi_N}{\partial \pi_N} \end{pmatrix} \quad \text{waar} \quad \frac{\partial \log \pi_i}{\partial \pi_i} = \frac{1}{\pi_i} \quad \& \quad \frac{\partial \log \pi_i}{\partial \pi_j} = \frac{0}{\pi_j} = 0 \quad \forall i \neq j \\ &= \begin{pmatrix} \frac{1}{\pi_1} & \dots & 0 \\ \dots & \ddots & \dots \\ 0 & \dots & \frac{1}{\pi_N} \end{pmatrix} \\ &= \text{Diag}(\underline{\pi})^{-1} \end{aligned}$$

Pas nou die deltametode toe sodat

$$\therefore \sqrt{n}(\log(\underline{p}) - \log(\underline{\pi})) \underset{*}{\sim} N \left(\underline{0}, \underbrace{\left(\frac{\partial \underline{g}}{\partial \underline{\pi}} \right) (D(\underline{\pi}) - \underline{\pi} \underline{\pi}') \left(\frac{\partial \underline{g}}{\partial \underline{\pi}} \right)'} \right)$$

$$\begin{aligned}
\text{Maar } * &= D(\underline{\pi})^{-1} (D(\underline{\pi}) - \underline{\pi}\underline{\pi}') D(\underline{\pi})^{-1} \\
&= (I - D(\underline{\pi})^{-1} \underline{\pi}\underline{\pi}') D(\underline{\pi})^{-1} \\
&= D(\underline{\pi})^{-1} - \underline{1} \underline{1}' \quad (\text{Kontroleer gerus hierdie stap deur die matrikse uit te skryf. Dis lomp maar dit werk.})
\end{aligned}$$

$$\therefore \sqrt{n}(\log(\underline{p}) - \log(\underline{\pi})) \sim N(\underline{0}, D(\underline{\pi})^{-1} - \underline{1} \underline{1}')$$

3. Pas nou 'n verskuiwing $C_{q \times N}$ van konstantes toe, waar $q = 1$.

$$\begin{aligned}
\therefore \sqrt{n}C(\log(\underline{p}) - \log(\underline{\pi})) &\sim N(\underline{0}, C(D(\underline{\pi})^{-1} - \underline{1} \underline{1}')C') \\
&\sim N(\underline{0}, CD(\underline{\pi})^{-1}C' - C\underline{1}\underline{1}'C')
\end{aligned}$$

Stel $C = [1 \ -1 \ 1 \ -1 \ 0 \ \dots 0]$, $q = 1$; $N = N$. Hierdie C word geskep op so 'n wyse dat die logkansverhoudingsfunksie daaruit soos volg na vore kom:

$$\begin{aligned}
\log \frac{p_1 p_3}{p_2 p_4} &= \log p_1 - \log p_2 + \log p_3 - \log p_4 \\
&= C(\log \underline{p})
\end{aligned}$$

$$\text{i.e.} \quad ([1 \ -1 \ 1 \ -1 \ 0 \ 0 \ \dots]) \begin{bmatrix} \log p_1 \\ \log p_2 \\ \log p_3 \\ \log p_4 \end{bmatrix}$$

$$\text{sodat } \text{cov}(\sqrt{n}C \log \underline{p}) = C(D(\underline{\pi})^{-1} - \underline{1} \underline{1}')C'$$

$$= CD(\underline{\pi})^{-1}C' - C\underline{1}\underline{1}'C'$$

$$= (1 \ -1 \ 1 \ -1) \begin{pmatrix} \frac{1}{\pi_1} & 0 & 0 & 0 \\ 0 & \frac{1}{\pi_2} & 0 & 0 \\ 0 & 0 & \frac{1}{\pi_3} & 0 \\ 0 & 0 & 0 & \frac{1}{\pi_4} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} - (1 \ -1 \ 1 \ -1) \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} (1 \ 1 \ 1 \ 1) \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{\pi_1} & \frac{-1}{\pi_2} & \frac{1}{\pi_3} & \frac{-1}{\pi_4} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} - 0 \cdot 0$$

$$= \frac{1}{\pi_1} + \frac{1}{\pi_2} + \frac{1}{\pi_3} + \frac{1}{\pi_4}$$

$$\begin{aligned} \therefore \quad \sqrt{n} \left(\log \frac{p_1 p_3}{p_2 p_4} - \log \frac{\pi_1 \pi_3}{\pi_2 \pi_4} \right) &\sim N \left(0, \frac{1}{\pi_1} + \frac{1}{\pi_2} + \frac{1}{\pi_3} + \frac{1}{\pi_4} \right) \\ \text{of} \quad \log \frac{p_1 p_3}{p_2 p_4} &\sim N \left(\log \frac{\pi_1 \pi_3}{\pi_2 \pi_4}, \frac{1}{n} \left(\frac{1}{\pi_1} + \frac{1}{\pi_2} + \frac{1}{\pi_3} + \frac{1}{\pi_4} \right) \right) \end{aligned}$$

Sodat

$$\ln \hat{\theta} \sim N \left(\ln \theta, \sum_i \sum_j \frac{1}{n_{ij}} \right)$$

Hiermee is die bewys dus voltooi.