Een Vergelijking tussen de Frequentistische en Bayesiaanse Benaderingen in de Statistiek

Aan de hand van het Duitse tank probleem

Auteur:
Sterre Manon Corver
6600948

Begeleider:
Dr. S. Dirksen

Universiteit Utrecht
Beta Faculteit
Wiskunde en Toepassingen

Abstract

Deze scriptie onderzoekt de verschillen tussen de frequentistische en Bayesiaanse benaderingen in de statistiek. De basisprincipes van de Bayesiaanse statistiek worden besproken en toegepast op het Duitse tank probleem, dat draait om het schatting van het aantal geproduceerde tanks op basis van de serienummers van in beslag genomen tanks. Voor de frequentistische benadering wordt een schatter afgeleid die de zuivere schatter met minimale variantie is. Er wordt aangetoond dat deze schatter inderdaad de zuivere schatter met minimale variantie is. In de Bayesiaanse benadering wordt het limietgedrag van de *posterior* kansfunctie geanalyseerd om een schatting te maken van het aantal tanks. Voor beide benaderingen worden de verwachtingswaarde en variantie afgeleid en gebruikt in de vergelijking van de twee benaderingen.

Inhoudsopgave

Hoofdstuk 1

Introductie

Een fundamenteel probleem binnen de statistiek is inferentie. Hierbij worden gegevens uit een steekproef geobserveerd en worden uitspraken (inferenties) gedaan over één of meer onbekende kenmerken van het onderliggende systeem dat deze gegevens genereert (de populatie) [14]. Er zijn twee belangrijke benaderingen binnen statistische inferentie: de frequentistische en de Bayesiaanse benadering [17]. Het verschil tussen deze benaderingen is dat in de frequentistische benadering de populatie parameters als vaste constanten worden beschouwd, terwijl deze in de Bayesiaanse benadering als kansvariabelen worden gezien. In deze scriptie worden de twee benaderingen beschreven en vergeleken aan de hand van het Duitse tank probleem. Er wordt gestart met een uitleg over Bayesiaanse statistiek en de bijbehorende regels. Enkele relevante basisprincipes van de frequentistische statistiek worden herhaald in Appendix A. Vervolgens worden zowel de frequentistische als de Bayesiaanse benaderingen van serienummer analyse toegepast op het Duitse tank probleem waarbij, met behulp van verwachtingswaarden, schatters worden afleidt en de varianties ervan worden bepaald. Aan de hand van deze uitkomsten worden in Hoofdstuk 5 de benaderingen vergeleken.

Serienummer analyse is een effectieve statistische methode om het totale aantal geproduceerde items in een reeks te schatten, gebaseerd op de serienummers van een steekproef van die items. Deze techniek werd beroemd door zijn toepassing tijdens de Tweede Wereldoorlog, bekend als het Duitse Tank Probleem, maar heeft sindsdien talloze toepassingen gevonden. Het Duitse tank probleem draait om de uitdaging om het totale aantal geproduceerde Duitse tanks te schatten aan de hand van de serienummers van buitgemaakte tanks. Deze methode vereist dat de serienummers in volgorde van productie zijn.

Naast de toepassing in militaire sferen, is er een voorbeeld van het gebruik van serienummer analyse door een Londense statisticus in 2008. Hij vroeg mensen om het serienummer en de aankoopdatum van hun iPhone naar hem te sturen om zo een schatting te maken van het totale aantal verkochte iPhones door Apple.
Om de methodologie te illustreren, stellen we ons een scenario voor waarin we langs de kant van de weg staan tijdens een marathon. Terwijl de lopers voorbij rennen, noteren we een aantal startnummers, bijvoorbeeld 104, 20, 331, 202 en 85. Op

basis van deze gegevens, ervan uitgaande dat de startnummers beginnen bij 1, is het duidelijk dat er minstens 331 deelnemers zijn, omdat dat het hoogste nummer is dat we hebben waargenomen. Dit geeft een ondergrens, maar voor een nauwkeurigere schatting kan er gebruik gemaakt worden van serienummer analyse.

Het Duitse tank probleem

Tijdens de Tweede Wereldoorlog probeerden de Westerse Geallieerden de omvang van de Duitse tank productie vast te stellen. Dit deden ze via verschillende methoden, zoals informatie verkregen door inlichtingenwerk, ondervragingen en spionage. Deze schattingen bleken echter onbetrouwbaar en onnauwkeurig [9]. Om deze reden werd ook statistiek ingezet, bijvoorbeeld door de serienummers van gevangen genomen tanks te analyseren.
In de tabel hieronder worden schattingen van de gemiddelde maandelijkse productie in 1940-1942 van het aantal tanks op basis van serienummer analyse vergeleken met de schattingen op basis van informatie ingewonnen door de inlichtingendienst. De schatter op basis van inlichtingenwerk is verkregen uit gecombineerde blikken van de Amerikaanse en Britse inlichtingendienst op dat moment. Deze staan beschreven in *Munitions Record No. 24*, welke op 10 augustus 1942 is gepubliceerd [5]. Omdat hier schattingen voor juni 1940, juni 1941 en augustus 1942 worden beschreven, zijn dit de maanden die gebruikt worden voor de vergelijking in de tabel.

Maand	**Daadwerkelijk aantal**	**Schatting aan de hand van serienummers**	**Schatting in *Munitions Record No. 24***
Juni 1940	122	169	1000
Juni 1941	271	244	1550
Augustus 1942	342	327	1550

Tabel 1.1: Tabel met (schattingen van) het aantal geproduceerde tanks per maand.

Volgens schattingen van de inlichtingendiensten produceerden de Duitsers dus ongeveer 1.370 tanks per maand tussen juni 1940 en augustus 1942. Door gebruik te maken van statistische schattingen werd het geschatte aantal vastgesteld op gemiddeld bijna 247 per maand. Na de oorlog zijn de daadwerkelijke productiecijfers geproduceerd, en bleek dat het daadwerkelijke aantal 245 was. De statistische schatting lag dus erg dicht bij het daadwerkelijke aantal.

Om de productiesnelheid van tanks in Duitsland te schatten met behulp van statistiek, verzamelden de geallieerden dus de serienummers van de gevangen genomen tanks. Als geallieerden weten we niet wat het totale aantal geproduceerde tanks n is, maar hebben we een steekproef k van ingenomen tanks. De informatie die we hieruit gebruiken zijn de serienummers $(s_1, ..., s_k)$. Ons doel is om het aantal geproduceerde tanks n, te schatten aan de hand van de serienummers. Hierbij nemen we aan dat elke geproduceerde tank even waarschijnlijk is om gevangen te worden genomen. Elke schatting van de n vanuit de gegevens $(s_1, ..., s_k)$ is onderhevig aan

onzekerheid, omdat we (waarschijnlijk) niet alle tanks hebben gevangengenomen. Om dit probleem zo overzichtelijk mogelijk te houden gaan we er vanuit dat de telling van de serienummers begint bij 1. Zodra een tank veroverd is dan blijft deze in bezit van de veroveraars, wat dus impliceert dat we te maken hebben met een kansexperiment zonder teruglegging.
Omdat maar twee kenmerken van de gegevens $(s_1, ..., s_k)$ informatie bevatten over de grootte van de tank populatie n, namelijk de omvang k en het maximale waargenomen serienummer m, wordt in het vervolg niet gebruik gemaakt van $(s_1, ..., s_k)$. In plaats daarvan worden k en m gebruikt, welke waardes volgen uit de serienummers van de in beslag genomen tanks [16].

Voor het overzicht volgen hieronder de definities van de variabelen die in de volgende hoofdstukken gebruikt worden:

- n = het totaal aantal tanks dat daadwerkelijk gefabriceerd is;
- k = het aantal veroverde tanks;
- m = het maximale serienummer onder de k veroverde tanks.

Hoofdstuk 2

Bayesiaanse statistiek

2.1 Introductie in Bayesiaanse statistiek

Bayesiaanse statistiek kun je als volgt zien: stel je voor dat je een schatkaart hebt waarop staat dat de schat ergens op een groot eiland ligt, maar je weet niet precies waar. Je kan de kaart en je observaties ter plaatse gebruiken om te bepalen waar je het beste op zoek kunt gaan naar de schat. Bayes' theorie geeft ons een formele manier om deze overgang van voorlopige aannames naar geïnformeerde overtuigingen te maken. Het draait dus allemaal om het idee van het aanpassen van onze initiële overtuigingen (de schatkaart), welke de *priors* worden genoemd, op basis van de beschikbare gegevens (de observaties ter plaatste), welke de *likelihoods* worden genoemd. Hieruit volgen dan de bijgewerkte overtuigingen, ofwel de *posterior* kansen. Bayesiaanse inferentie is dus een proces van leren van data [13]. Dit proces wordt beschreven door de Stelling van Bayes. Deze stelling zegt dat

$$P(B_i|A) = \frac{P(A|B_i) \times P(B_i)}{P(A)} \text{ [6].}$$

De kern van Bayes' theorie is eenvoudig: de kans op een gebeurtenis na het verkrijgen van nieuwe informatie is evenredig met de kans voor die informatie, vermenigvuldigd met de kans op die informatie gegeven de gebeurtenis. Dit principe, vaak samengevat als $\textit{posterior} \propto \textit{prior} \times \textit{likelihood}$, vormt de basis van Bayesiaanse statistiek.

Deze stelling van Bayes komt voort uit de definitie van voorwaardelijke kans. De marginale kans van gebeurtenis A wordt gevonden door de wet van totale kans, en elk van de gezamenlijke kansen wordt gevonden met behulp van de vermenigvuldigingsregel (Appendix A). $P(B_i)$ wordt de *prior* kans genoemd van de gebeurtenis B_i, $P(A)$ is de geobserveerde gebeurtenis, ofwel de *likelihood*. $P(B_i|A)$ wordt de *posterior* kans van gebeurtenis B_i genoemd.

2.2 Discrete kansvariabelen in Bayesiaanse statistiek

Een kansvariabele Y wordt als discreet beschouwd als deze waarden $y_1, y_2, ..., y_k$ aanneemt in een aftelbare deelverzameling van $\mathbb{R}$. De kansfunctie van een discrete kansvariabele geeft de kans geassocieerd met elke mogelijke waarde, dus $p_Y(y_i) = P(Y = y_i)$.
In de Bayesiaanse benadering wordt de parameter die we willen schatten als een kansvariabele X, met mogelijke waarden $x_1, x_2, ..., x_I$, beschouwd. De geobserveerde kansvariabele Y, die afhankelijk is van de parameter, heeft mogelijke waarden $y_1, y_2, ..., y_J$ [6].

De *Prior* verdeling

De *prior* verdeling van de discrete kansvariabele X wordt gegeven door de *prior* kansfunctie $p_X(x_i) = P(X = x_i)$, voor $i = 1, 2, ..., I$. Dit vertegenwoordigt onze initiële geloofwaardigheid in de kans van elke mogelijke waarde x_i, voordat we naar de gegevens kijken. Het moet voortkomen uit eerdere ervaringen of overtuigingen, en dus niet uit de huidige gegevens.

De *Likelihood*

De functie $P(Y = y_j|X = x_i)$, waarbij $i = 1, ..., I$ en $j \in \{1, ..., J\}$, beschrijft de *likelihood* in het geval van discrete kansvariabelen. Dit is de voorwaardelijke kansfunctie van Y gegeven dat $X = x_i$ wordt geëvalueerd op y_j, de daadwerkelijke geobserveerde waarde. Hierbij varieert X over zijn hele bereik $x_1, ..., x_n$. We moeten weten wat voor verdeling de geobserveerde variabele heeft, omdat dit laat zien hoe de verdeling van de observatie Y afhankelijk is van de waarde van de kansvariabele X, maar we hoeven deze alleen te evalueren op de daadwerkelijk geobserveerde waarde, y_j. De *likelihood* functie is dus de voorwaardelijke kansverdeling van de observatie geëvalueerd op het gereduceerde universum $Y = y_j$.

De *Posterior* verdeling

De *posterior* verdeling van X gegeven $Y = y_j$ wordt gegeven door de kansfunctie

$$P(X = x_i|Y = y_j) = \frac{P(X = x_i) \times P(Y = y_j|X = x_i)}{\sum_{i=1}^{I} P(X = x_i) \times P(Y = y_j|X = x_i)} \text{ voor } i = 1, \ldots, I.$$

Deze functie geeft een methode om de geloofwaardigheid over de mogelijke waarden van X te herzien, gegeven dat $Y = y_j$ is waargenomen.

Hoofdstuk 3

Frequentistisch Duitse tank probleem

In de frequentistische statistiek is de Meest Aannemelijke Schatter methode de meest populaire techniek voor het afleiden van schatters [15]. Voor het Duitse tank probleem is het echter niet mogelijk om gebruik te maken van deze schatter voor het bepalen van het aantal tanks n. Deze schatter stelt namelijk dat de schatting van het aantal tanks n gelijk is aan het maximale serienummer m onder de k gevangen genomen tanks. Deze is duidelijk onzuiver omdat $m \leq n$ [16].
In de frequentistische benadering van het Duitse tank probleem wordt daarom gebruik gemaakt van een andere schatter. Dit is de Zuivere Schatter met Minimale Variantie, welke is afgeleid door Leo Goodman [7]. De Zuivere Schatter met Minimale Variantie geeft op de volgende manier een schatting voor het aantal tanks in de populatie: de schatter stelt dat het verwachte aantal geproduceerde tanks gelijk is aan het hoogste serienummer in de steekproef plus het gemiddelde aantal wat tussen de gevangen genomen tanks zit, dus $\hat{n} = m + \frac{m}{k} - 1$. De afleiding van deze schatter volgt hieronder.

3.1 Kansfunctie

Om de verwachting van het aantal tanks te kunnen bepalen moet eerst de bijbehorende kansfunctie gedefinieerd worden. De kansfunctie die de kans dat de maximale steekproef waarde M gelijk is aan m beschrijft, wordt genoteerd met $P(M = m)$. Voor elke m kan de kans dat het de maximale waarde is, worden bepaald. Omdat het niet mogelijk dat het maximale serienummer kleiner is dan het aantal in beslag genomen tanks, geldt voor $m < k$ dat $P(M = m) = 0$. Voor de m waarvoor geldt dat $k \leq m \leq n$, wordt aangenomen dat het een uniforme verdeling, in een experiment zonder teruglegging, volgt. De kansfunctie van M bestaat uit het aantal manieren om $(k-1)$ tanks te selecteren uit een verzameling van $(m-1)$ tanks, gedeeld door het aantal manieren om k tanks te selecteren uit de totale verzameling van n tanks. Om dit in formule vorm te schrijven kan er gebruik gemaakt worden van de binomiaal coëfficiënt. De kansfunctie voor $m = k, \ldots, n$ wordt dan gegeven

door

$$P(M = m) = \frac{\binom{m-1}{k-1}}{\binom{n}{k}} \text{ [12].} \tag{3.1}$$

3.2 Puntschatter

Om een punt schatter voor het aantal tanks n te bepalen, wordt er begonnen met het bepalen van de verwachtingswaarde van M. Hiertoe wordt de functie voor de verwachtingswaarde (5.1) in appendix A toepast op het Duitse tank probleem en de daarvoor gespecificeerde variabelen:

$$E(M) = \sum_{m=1}^{n} mP(M = m).$$

Aangezien de kansfunctie verschillend is voor $m < k$ en $k \leq m \leq n$ wordt de sommatie in twee delen opgesplitst. Omdat $P(M = m) = 0$ voor $m < k$ geldt dat

$$E(M) = \sum_{m=1}^{k-1} mP(M = m) + \sum_{m=k}^{n} mP(M = m) = \sum_{m=k}^{n} mP(M = m),$$

en als hier de kansfunctie van m voor $k \leq m \leq n$ wordt ingevuld dan volgt hieruit dat

$$E(M) = \sum_{m=k}^{n} m\frac{\binom{m-1}{k-1}}{\binom{n}{k}}.$$

Om hieruit een schatting van n te kunnen maken moet de functie vereenvoudigd worden. Als eerste wordt er gebruik gemaakt van de uitsplitsing van de binomiaal coëfficiënt [6]. De binomiaal coëfficiënt is zo gedefinieerd dat

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

De uitsplitsing toegepast op $\binom{m-1}{k-1}$ geeft dat

$$E(M) = \sum_{m=k}^{n} m\frac{\binom{m-1}{k-1}}{\binom{n}{k}} = \sum_{m=k}^{n} m\frac{\frac{(m-1)!}{(k-1)!((m-1)-(k-1))!}}{\binom{n}{k}} = \sum_{m=k}^{n} \frac{\frac{m!}{(k-1)!(m-k)!}}{\binom{n}{k}},$$

en door de functie $\frac{m!}{(k-1)!(m-k)!}$ met $\frac{k}{k}$ te vermenigvuldigen kan het terug worden gezet in een binomiaal coëfficiënt. We krijgen

$$E(M) = \sum_{m=k}^{n} \frac{\frac{km!}{k!(m-k)!}}{\binom{n}{k}} = \sum_{m=k}^{n} \frac{k\binom{m}{k}}{\binom{n}{k}}.$$

Aangezien k en $\binom{n}{k}$ niet worden beïnvloed door de sommatie kunnen deze uit de sommatie worden gehaald, en vereenvoudigd de verwachting naar

$$E(M) = \frac{k}{\binom{n}{k}} \sum_{m=k}^{n} \binom{m}{k}.$$

De Wet van Totale Kans stelt dat voor een discrete kansvariabele de som van de kansen altijd gelijk moet zijn aan 1. Door gebruik te maken van de eerder gevonden kansfunctie, en het feit dat $\binom{n}{k}$ niet wordt beïnvloed wordt door de sommatie, kan de sommatie worden weggewerkt. We krijgen

$$\sum_{m=k}^{n} P(M=m) = \sum_{m=k}^{n} \frac{\binom{m-1}{k-1}}{\binom{n}{k}} = \frac{1}{\binom{n}{k}} \sum_{m=k}^{n} \binom{m-1}{k-1} = 1, \text{ dus}$$

$$\binom{n}{k} = \sum_{m=k}^{n} \binom{m-1}{k-1}. \tag{3.2}$$

De formule van de verwachtingswaarde bevat $\binom{m}{k}$ in plaats van $\binom{m-1}{k-1}$. Toch kan deze uitkomst gebruikt worden door de sommatie te verschuiven van $m = k$ tot n naar $m = k + 1$ tot $n + 1$. De verwachting wordt dan

$$E(M) = \frac{k}{\binom{n}{k}} \binom{n+1}{k+1}.$$

Om dit te vereenvoudigen schrijven we de binomiaal coëfficiënten weer uit:

$$\begin{aligned} E(M) &= \frac{k(n+1)!}{((n+1)-(k+1))!(k+1)!} \frac{(n-k)!k!}{n!} \\ &= \frac{k(n+1)!}{(n-k)!(k+1)!} \frac{(n-k)!k!}{n!} \end{aligned}$$

Omdat $(n-k)!$ zowel in de teller als in de noemer voorkomt, valt dit weg. Ook geldt dat $\frac{k!}{(k+1)!} = \frac{1}{k+1}$ en $\frac{(n+1)!}{n!} = n + 1$. Er volgt dat

$$E(M) = \frac{k(n+1)}{k+1}.$$

Het doel is om een schatting te maken van het aantal tanks n. Door de vergelijking van $E(M)$ als volgt herschikken, kan een zuivere punt schatter voor n worden afgeleid:

$$\begin{aligned} \frac{k(n+1)}{k+1} &= E(M) \\ kn + k &= E(M)(k+1) \\ kn &= E(M)(k+1) - k \\ n &= E(M) + \frac{E(M)}{k} - 1. \end{aligned}$$

Als de schatter wordt gedefinieerd door,

$$\hat{n} = M + \frac{M}{k} - 1 \tag{3.3}$$

dan geldt dat

$$E(\hat{n}) = E(M) + \frac{E(M)}{k} - 1 = n,$$

en het dus een zuivere schatter is.

3.3 Variantie van de puntschatter

Om de variantie van $\hat{n}$ te bepalen wordt gebruik gemaakt de eigenschap van de variantie dat $Var[aY + b] = a^2 Var[Y]$. Dit toegepast op het Duitse tank probleem geeft

$$\begin{aligned} Var(\hat{n}) &= Var\left(M + \frac{M}{k} - 1\right) && (3.4)\\ &= Var\left(M\frac{k+1}{k} - 1\right) && (3.5)\\ &= \left(\frac{k+1}{k}\right)^2 \cdot Var(M). && (3.6) \end{aligned}$$

Om de variantie van $\hat{n}$ te vinden moet eerst de variantie van M worden bepaald. De functie voor variantie in Appendix A (5.2) geeft dat

$$Var(M) = E(M^2) - E(M)^2 = E(M(M+1)) - E(M) - E(M)^2,$$

omdat $E(M(M+1)) - E(M) = E(M^2) + E(M) - E(M) = E(M^2)$.

Voor het Duitse tank probleem geldt dat

$$\begin{aligned} E(M(M+1)) &= \sum_{m=k}^{n} m(m+1)P(M=m) \\ &= \sum_{m=k}^{n} m(m+1)\frac{\binom{m-1}{k-1}}{\binom{n}{k}}. \end{aligned}$$

Voor natuurlijke getallen m, k met $k \leq m$ geldt dat $\binom{m}{k} = \frac{m}{k}\binom{m-1}{k-1}$ voor $k \geq 1$. Dus $\frac{m(m+1)}{k(k+1)}\binom{m-1}{k-1} = \binom{m+1}{k+1}$. Dit geeft

$$\begin{aligned} E(M(M+1)) &= \sum_{m=k}^{n} m(m+1)\frac{\binom{m-1}{k-1}}{\binom{n}{k}} \\ &= \sum_{m=k}^{n} k(k+1)\frac{m(m+1)}{k(k+1)}\frac{\binom{m-1}{k-1}}{\binom{n}{k}} \\ &= \frac{k(k+1)}{\binom{n}{k}} \sum_{m=k}^{n} \binom{m+1}{k+1}, \end{aligned}$$

waarbij $\sum_{m=k}^{n} \binom{m-1}{k-1} = \binom{n}{k}$ (3.2), en dus $\sum_{m=k}^{n} \binom{m+1}{k+1} = \binom{n+2}{k+2}$ geeft

$$
\begin{aligned}
&= \frac{k(k+1)}{\binom{n}{k}} \binom{n+2}{k+2} \\
&= \frac{k(k+1)}{\frac{n!}{k!(n-k)!}} \frac{(n+2)!}{(k+2)!((n+2)-(k+2))!} \\
&= \frac{k(k+1)k!(n-k)!}{n!} \frac{(n+2)!}{(k+2)!(n-k)!} \\
&= \frac{k(k+1)!}{n!} \frac{(n+2)!}{(k+2)!} \\
&= \frac{k(k+1)!(n+2)!}{n!(k+2)!} \\
&= \frac{k(n+2)(n+1)}{(k+2)}
\end{aligned}
$$

Nu $E(M(M+1))$ is bepaald, kan er verder worden gegaan met het bepalen van de variantie van M:

$$
\begin{aligned}
Var(M) &= E(M(M+1)) - E(M) - E(M)^2 \\
&= \frac{k(n+2)(n+1)}{(k+2)} - \frac{k(n+1)}{k+1} - \frac{k^2(n+1)^2}{(k+1)^2} \\
&= \frac{k(n+2)(n+1)}{(k+2)} - \frac{k(n+1)}{k+1} \left(1 + \frac{k(n+1)}{k+1}\right) \\
&= \frac{k(n+2)(n+1)}{(k+2)} - \frac{k(n+1)}{k+1} \left(\frac{(k+1)+k(n+1)}{k+1}\right) \\
&= k(n+1) \left(\frac{(n+2)}{(k+2)} - \frac{1}{k+1} \frac{(k+1)+k(n+1)}{k+1}\right) \\
&= k(n+1) \left(\frac{(n+2)}{(k+2)} - \frac{(k+1)+k(n+1)}{(k+1)^2}\right) \\
&= k(n+1) \left(\frac{(n+2)}{(k+2)} - \frac{(1+k(n+2)}{(k+1)^2}\right) \\
&= k(n+1) \left(\frac{(n+2)(k+1)^2}{(k+2)(k+1)^2} - \frac{(1+k(n+2))(k+2)}{(k+1)^2(k+2)}\right) \\
&= k(n+1) \frac{(n+2)(k+1)^2 - ((1+k(n+2))(k+2)}{(k+2)(k+1)^2} \\
&= k(n+1) \frac{(n+2)(k^2+2k+1) - (n+2)(k^2+2k) - (k+2)}{(k+2)(k+1)^2} \\
&= k(n+1) \frac{(n+2) - (k+2)}{(k+2)(k+1)^2} \\
&= \frac{k(n+1)(n-k)}{(k+2)(k+1)^2}.
\end{aligned}
$$

Met de gevonden variantie van M kan de formule voor $Var(\hat{n})$ (3.6) verder worden uitgewerkt. We krijgen

$$\begin{aligned} Var(\hat{n}) &= \left(\frac{k+1}{k}\right)^2 \cdot Var(M) \\ &= \frac{(k+1)^2}{k^2} \cdot \frac{k(n+1)(n-k)}{(k+2)(k+1)^2}, \\ &= \frac{(k+1)^2 k(n+1)(n-k)}{k^2(k+2)(k+1)^2} \\ &= \frac{(n-k)(n+1)}{k(k+2)}. \end{aligned}$$

3.4 Bewijs Zuivere Schatter met Minimale Variantie

Hieronder volgt een bewijs van dat schatter van n, $\hat{n} = M + \frac{M}{k} - 1$ (3.2), inderdaad de schatter met minimale variantie is.
Er geldt dat M een voldoende statistiek is voor n. Een statistiek $T(S)$ wordt voldoende genoemd voor θ als er geen andere statistiek bestaat die meer informatie bevat over de waarde van θ. Formeel gezien, neem aan dat $S = (S_1, \ldots, S_k)$ een onafhankelijke en identiek verdeelde willekeurige steekproef is met S_i P_θ. Een statistiek $T(S)$ wordt voldoende voldoende genoemd voor een parameter θ als $E_\theta(h(S)|T(S))$ niet afhankelijk van is van θ elke meetbare functie h [11]. De factorisatie stelling van Neyman-Fisher zegt dat een statistiek $T(S)$ voldoende is voor θ dan en slechts dan als de gezamenlijke kansverdeling van de steekproef $S = (S_1, \ldots, S_k)$ kan worden geschreven als:

$$f(s_1 \ldots, s_k; \theta) = h(s_1, \ldots, s_k) \cdot g(T(s); \theta)$$

waarbij

- $h(s_1, \ldots, s_k)$ een functie is die niet afhangt van θ.
- $g(T(s); \theta)$ een functie die allen via de voldoende statistiek $T(S)$ en de parameter θ afhankelijk is van de data [11].

Om te voldoen aan de factorisatie stelling, moeten we de gezamenlijke kansverdeling dus uitdrukken als een product van twee functies: één die afhangt van de data via de voldoende statistiek $T(s)(= m)$ en n, en één die niet afhangt van n. Omdat k een vast gegeven is in de kansfunctie en dus geen parameter die moet worden geschat, is de aanwezigheid van k in de uitdrukking geen probleem voor het bewijs.
Formule (3.1) kan met behulp van de indicator functie $\mathbb{1}_{\{k \leq m \leq n\}}$ geschreven worden als

$$P(M = m) = \frac{\binom{m-1}{k-1}}{\binom{n}{k}} \mathbb{1}_{\{k \leq m \leq n\}} = \frac{\binom{m-1}{k-1}}{\binom{n}{k}} \mathbb{1}_{\{m \geq k\}} \mathbb{1}_{\{m \leq n\}}.$$

Hier hebben we:

- $g(m,n) = \frac{\binom{m-1}{k-1}}{\binom{n}{k}} \mathbb{1}_{\{m \leq n\}}$, die afhankelijk is van m en n;
- $h(m) = \mathbb{1}_{\{m \geq k\}}$, die niet afhankelijk is van n.

Volgens de factorisatie stelling hebben we dus aangetoond dat

$$P(M = m) = h(m) \cdot g(m; n),$$

waarbij m de grootste waarde in de steekproef $(s_1, ..., s_k)$ is. Dit toont aan dat M een voldoende statistiek is voor n.

Omdat M een voldoende statistiek is, kan met behulp van de Rao-Blackwell stelling worden aangetoond dat er alleen rekening gehouden hoeft te worden met met schatters voor n die een functie zijn van M.
Laat $T_k(S) = T(S_1, ..., S_k)$ een voldoende statistiek statistiek zijn en θ^* een schatter. Definieer $\theta^*_S = E_\theta(\theta^*|T_k(S))$ a.s.. Dan zegt de stelling dat voor alle $\theta \in \Theta$ we hebben dat $E_\theta[(\theta^*_s - \theta)^2] \leq E_\theta[(\theta^* - \theta)^2]$.
Het bewijs van deze stelling volgt uit de ongelijkheid van Jensen. We hebben

$$\begin{aligned}(\theta^*_S - \theta)^2 &= (E_\theta(\theta^*|T_k(S)) - \theta)^2 \\ &= (E_\theta(\theta^* - \theta|T_k(S)))^2 \\ &\leq E_\theta((\theta^* - \theta)^2|T_k(S))\end{aligned}$$

Als aan beide kanten de verwachting wordt genomen, en er gebruik gemaakt wordt van het feit dat voor $y < \infty$, $E(E(y|x)) = E(y)$ [2], geldt er

$$E_\theta[(\theta^*_S - \theta)^2] \leq E_\theta[E_\theta((\theta^* - \theta)^2|T_k(S))] = E_\theta[(\theta^* - \theta)^2].$$

Dus de variantie van de conditionele verwachting van y gegeven x is altijd kleiner dan of gelijk aan de variantie van y. Dit betekent dat door gebruik te maken van de conditionele verwachting als schatter, de variantie kan worden verminderd.
Deze resultaten impliceren dat als er een voldoende statistiek $T(S)$ is, en daarnaast een zuivere schatter $\hat{\theta}$ voor een parameter θ met $\hat{\theta}$ een functie van zowel $T(S)$ als andere variabelen, de voorwaardelijke verwachting $E(\hat{\theta}|T(S))$ een zuivere schatter is voor θ. Bovendien heeft $E(\hat{\theta}|T(S))$ een variantie die kleiner is dan of gelijk aan die van $\hat{\theta}$.

Door gebruik te maken van de stelling van Rao-Blackwell, kan worden geconcludeerd dat er alleen gekeken hoeft te worden naar schatters die functies zijn van M, omdat M een voldoende statistiek is voor n. Er geldt dat alleen zuivere schatters kunnen worden gevonden die alleen afhankelijk zijn van M met vergelijkbare of betere varianties dan die van schatters die afhankelijk zijn van andere variabelen.

Nu volgt het bewijs waarom er maar één functie $e(m)$ bestaat welke een zuivere voldoende schatter is en dat die gelijk moet zijn aan $\hat{n}$. Voor een gegeven steekproef grootte van k zijn de mogelijke waardes voor de parameter n gelijk aan $(k, k+1, k+2, \ldots)$. Daarnaast zijn de mogelijke waarden voor het grootste serienummer in de

steekproef gelijk aan $m = k, k+1, k+2, \ldots$. Als er geldt dat $n = k$, dan moet ook gelden dat $m = k$. Dus de kans dat $m = k$ gegeven $n = k$ is gelijk aan 1: $P(m = k|n = k) = 1$. Dus als we willen dat $e(m)$ een zuivere schatter is van n, moet gelden dat $e(m) = k$ wanneer $n = k$. Dus dat

$$E(e(m)|n = k) = e(k) = k.$$

Als $n = k+1$, dan zijn de enige mogelijke waardes voor m: $m = k$ en $m = k+1$. Dit zijn dus de enige waarden waarvoor de kans groter is dan 0. Dus, als $e(m)$ zuiver is voor alle waardes van n moet gelden dat $e(m) = k+1$ als $n = k+1$. Daarom geldt dat

$$\begin{aligned} E(e(m)|n = k+1) &= e(k)P(m = k|n = k+1) + e(k+1)P(m = k+1|n = k+1) \\ &= kP(m = k|n = k+1) + e(k+1)P(m = k+1|n = k+1) \end{aligned}$$

gelijk moet zijn aan $k+1$. Aangezien de waardes van de kansen berekend kunnen worden kan de vergelijking worden herschreven als

$$\begin{aligned} e(k+1)P(m = k+1|n = k+1) &= k+1-kP(m = k|n = k+1), \text{ en dus} \\ e(k+1) &= \frac{k+1-kP(m = k|n = k+1)}{P(m = k+1|n = k+1)}, \end{aligned}$$

wat de waarde geeft waar $e(k+1)$ gelijk aan moet zijn om te zorgen dat $e(m)$ een zuivere schatter is.

Op dezelfde manier geldt voor $n = k+2$ dat

$$e(k+2) = \frac{k+2-kP(m = k|n = k+2) - e(k+1)P(m = k+1|n = k+2)}{P(m = k+2|n = k+2),}$$

en omdat de getallen aan de rechterkant van de vergelijking berekend kunnen worden, kan $e(k+2)$ berekend worden en is het dus uniek.

Algemener valt te zien dat als $e(m)$ zuiver is,

$$e(h) = \frac{h - \sum_{i=k}^{h-1} e(i)P(m = i|n = h)}{P(m = h|n = h)}, \text{voor } h = k+1, \ldots, n, \text{ en } e(h) = k \text{ voor } h = k.$$

De recursieve relatie voor de waardes van $e(h)$ bepaalt $e(h)$ uniek, en daarom is $e(m)$ de enige zuivere schatter van n welke alleen op m gebaseerd is. Dus geldt dat $e(m) = \hat{n}$, en is bewezen dat de schatter (3.3) die hebben afgeleid in Hoofdstuk 3.2 de Zuivere Schatter met Minimale Variantie is.

Hoofdstuk 4

Bayesiaans Duitse tank probleem

In de Bayesiaanse benadering wordt het aantal tanks n beschouwd als een kansvariabele. Er wordt begonnen met een overtuiging van de verdeling van n welke wordt beschreven door de *prior* kansfunctie. Daarna worden deze overtuigingen bijgewerkt aan de hand van de waargenomen serienummers $(s_1, ..., s_k)$. Dit noemen we de waarschijnlijkheden (*likelihoods*). Met behulp van deze twee functies kan dan de *posterior* kansfunctie worden gevonden. Deze functie bevat de informatie bevat die gebruikt kan worden om het aantal tanks te schatten. De *posterior* kansfunctie van N kent aan elke mogelijke tank populatiegrootte n een waarschijnlijkheid toe op basis van een compromis tussen de *likelihoods* en de *prior* kansfunctie [16].

4.1 Kansfunctie

De *posterior* kansfunctie is uit Hoofdstuk 2, toegepast op het Duitse tank probleem geeft

$$P(N = n|M = m, K = k) = \frac{P(M = m, K = k|N = n) \cdot P(N = n)}{\sum_{i=0}^{n_{max}} P(M = m, K = k|N = i) \cdot P(N = i)}.$$

De *prior* kansfunctie

Stel dat er een bovengrens is voor het aantal tanks dat geproduceerd kan zijn: n_{max}. Dit zou bijvoorbeeld gebaseerd kunnen zijn op de beschikbaarheid van grondstoffen die nodig zijn voor tank productie, maar verder geen andere informatie hebben, dan kunnen we een diffuse prior gebruiken.
De *prior* kansfunctie beschrijft dat het verachte totale aantal tanks n elke waarde kan aannemen in $0, ..., n_{max}$, waarbij elke waarde even waarschijnlijk is. De kansfunctie die daarbij hoort is

$$P(N = n) = \frac{1}{n_{max} + 1} \text{ voor } 0 \leq n \leq n_{max} \text{ en } 0 \text{ anders [16].}$$

De *likelihood*

De likelihood is gelijk aan de frequentistische kansfunctie (3.1), dus

$$P(M = m, K = k|N = n) = \frac{\binom{m-1}{k-1}}{\binom{n}{k}} \text{ voor } k \leq m \leq n \text{ en 0 anders.}$$

De *posterior* kansfunctie

Hieruit volgt dat de *posterior* kansfunctie voor $k \leq m \leq n_{max}$ gelijk is aan

$$P(N = n|M = m, K = k) = \frac{\frac{\binom{m-1}{k-1}}{\binom{n}{k}} \times \frac{1}{n_{max}+1}}{\sum_{i=m}^{n_{max}} \frac{\binom{m-1}{k-1}}{\binom{i}{k}} \times \frac{1}{n_{max}+1}} = \frac{\binom{n}{k}^{-1}}{\sum_{i=m}^{n_{max}} \binom{i}{k}^{-1}}.$$

Voor het vereenvoudigen van de kansfunctie is het vereist om de sommatie wegwerken. Dit kan gedaan worden door de limiet van $n_{max} \to \infty$ te nemen, zodat de bovengrens van de sommatie ∞ wordt:

$$P(N = n|M = m, K = k) = \frac{\binom{n}{k}^{-1}}{\sum_{i=m}^{\infty} \binom{i}{k}^{-1}}.$$

Door de limiet naar oneindig te nemen wordt de *prior* echter wel improper. De *posterior* verdeling met de impropere *prior* verdeling, wordt verkregen door de limiet naar oneindig te nemen van de *posterior* verdeling met de propere *prior* [1]. Dit heeft negatieve gevolgen voor de schatter, maar door deze simplificatie is het mogelijk om gebruik te maken van hypergeometrische functies om de sommatie weg te werken. De hypergeometrische functie kan worden beschouwd als een genererende analytische functie van meer complexe combinatoriële getallen, die de binomiaal-reeks generaliseert [10]. Er zijn verschillende resultaten over hypergeometrische functies, waarvan de belangrijkste met betrekking tot ons probleem de hypergeometrische functie van Gauss is. Deze is gedefinieerd door

$${}_2F_1[a, b, c; 1] = \sum_{i=0}^{\infty} \frac{(a)^i (b)^i}{(c)^i} \frac{1}{i!} = \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)},$$

waarbij $(x)^i$ staat voor de stijgende faculteit [8]. De stijgende faculteit is als volgt gedefinieerd:

$$(x)^i = \frac{(x+i-1)!}{(x-1)!} = \frac{\Gamma(x+i)}{x} = (x+i-1) \cdot ... \cdot x.$$

In zowel definitie van de hypergeometrische functie als die van de stijgende faculteit komt de Gamma functie voor. De Gamma functie is een continue generalisering van de faculteit functie. De Gamma functie is zo gedefinieerd dat $\Gamma(n+1) = n!$ [3].

Als dit wordt toegepast op het Duitse tank probleem dan volgt

$$\sum_{i=m}^{\infty} \binom{i}{k}^{-1} = \sum_{j=0}^{\infty} \binom{m+j}{k}^{-1}$$

$$
\begin{aligned}
&= \sum_{j=0}^{\infty} \frac{k!(m+j-k)!}{(m+j)!} \\
&= \sum_{j=0}^{\infty} \frac{k!(m-k)!}{m!} \cdot \frac{m!}{(m+j)!} \cdot \frac{(m+j-k)!}{(m-k)!} \text{ (vermenigvuldig met } \frac{m!}{m!} \text{ en } \frac{(m-k)!}{(m-k)!}\text{)} \\
&= \binom{m}{k}^{-1} \sum_{j=0}^{\infty} \frac{m!}{(m+j)!} \cdot \frac{(m+j-k)!}{(m-k)!}.
\end{aligned}
$$

De definitie van de stijgende faculteit impliceert dat

$$
\begin{aligned}
\frac{m!}{(m+j)!} &= (\frac{(m+j)!}{m!})^{-1} = ((m+1)^j)^{-1} \text{ en} \\
\frac{(m+j-k)!}{(m-k)!} &= \frac{((m-k)+j)!}{(m-k)!} = (m-k+1)^j.
\end{aligned}
$$

Met behulp van de hypergeometrische functie van Gauss, kan de sommatie weggewerkt worden:

$$
\begin{aligned}
\sum_{i=m}^{\infty} \binom{i}{k}^{-1} &= \binom{m}{k} \sum_{j=0}^{\infty} \frac{(m-k+1)^j}{(m+1)^j} \\
&= \binom{m}{k}^{-1} \sum_{j=0}^{\infty} \frac{(m-k+1)^j}{(m+1)^j} \frac{(1)^j}{j!} \text{ want } (i)^j = j! \\
&= \binom{m}{k}^{-1} \sum_{j=0}^{\infty} \frac{(m-k+1)^j * (1)^j}{(m+1)^j} \frac{1}{j!} \\
&= \binom{m}{k}^{-1} \sum_{j=0}^{\infty} \frac{(m-k+1)^j * (1)^j}{(m+1)^j} \frac{1}{j!} \\
&= \binom{m}{k}^{-1} {}_2F_1[m-k+1, 1, m+1; 1] \\
&= \binom{m}{k}^{-1} \frac{\Gamma(m+1)\Gamma(m+1-m+k-1-1)}{\Gamma(m+1-m+k-1)\Gamma(m+1-1)} \\
&= \binom{m}{k}^{-1} \frac{\Gamma(m+1)\Gamma(k-1)}{\Gamma(k)\Gamma(m)}.
\end{aligned}
$$

Voor de Gamma functie geldt dat $\Gamma(n+1) = n!$, wat impliceert dat $\Gamma(n) = (n-1)!$ en $\Gamma(n-1) = (n-2)!$. Hieruit volgt dat

$$
\sum_{i=m}^{\infty} \binom{i}{k}^{-1} = \binom{m}{k}^{-1} \frac{m!(k-2)!}{(k-1)!(m-1)!}
$$

$$= \binom{m}{k}^{-1} \frac{m}{k-1}.$$

De gevonden benadering voor de sommatie kan ingevuld worden in de *posterior* kansfunctie:

$$P(N = n|M = m, K = k) = \frac{\binom{n}{k}^{-1}}{\frac{m}{k-1}\binom{m}{k}^{-1}} = \frac{(k-1)}{m}\frac{\binom{m}{k}}{\binom{n}{k}}.$$

4.2 Verwachtingswaarde

Nu er een benadering voor de *posterior* kansfunctie is gevonden, kan de verwachtingswaarde voor het aantal geproduceerde tanks opgesteld worden. Hierbij moet wel rekening worden gehouden met dat het de kansfunctie voor $n_{max} \to \infty$ is. De verwachtingswaarde die hierbij hoort is

$$\begin{aligned} E(N|M = m, K = k) &= \sum_{i=m}^{\infty} i \cdot P(N = i|M = m, K = k) \\ &= \sum_{i=m}^{\infty} i \cdot \frac{k-1}{m} \cdot \frac{\binom{m}{k}}{\binom{i}{k}} \\ &= \frac{k-1}{m} \cdot \binom{m}{k} \cdot \sum_{i=m}^{\infty} \frac{i}{\binom{i}{k}} \end{aligned}$$

Om dit te vereenvoudigen wordt begonnen met het evalueren van $\sum_{i=m}^{\infty} \frac{i}{\binom{i}{k}}$. Het uitschrijven van $\frac{i}{\binom{i}{k}}$ geeft

$$\frac{i}{\binom{i}{k}} = \frac{i}{\frac{i!}{k!(i-k)!}} = \frac{ik!(i-k)!}{i!} = \frac{k!(i-k)!}{(i-1)!} = \frac{k(k-1)!(i-k)!}{(i-1)!} = k\binom{i-1}{k-1}^{-1}.$$

Nu wordt wederom gebruik gemaakt van de identiteit $\sum_{i=m}^{\infty}\binom{i}{k}^{-1} = \frac{m}{k-1}\binom{m}{k}^{-1}$, alleen met $i-1$ in plaats van i en $k-1$ in plaats van k:

$$\sum_{i=m}^{\infty}\binom{i-1}{k-1}^{-1} = \frac{m-1}{k-2}\binom{m-1}{k-1}^{-1}.$$

De formule voor de verwachtingswaarde wordt

$$E(N|M = m, K = k) = \frac{k-1}{m} \cdot \binom{m}{k} \cdot k \cdot \frac{m-1}{k-2} \cdot \binom{m-1}{k-1}^{-1},$$

waarbij

$$\frac{\binom{m}{k}}{\binom{m-1}{k-1}} = \frac{\frac{m!}{k!(m-k)!}}{\frac{(m-1)!}{(k-1)!((m-1)-(k-1))!}}$$

$$= \frac{m!(k-1)!(m-k)!}{(m-1)!k!(m-k)!}$$
$$= \frac{m}{k}.$$

De verwachtingswaarde van het aantal geproduceerde tanks in de Bayesiaanse benadering is dus

$$E(N|M=m,K=k) = \frac{(k-1)\cdot(m-1)\cdot k\cdot m}{m\cdot(k-2)\cdot k} = (m-1)\frac{k-1}{k-2}.$$

4.3 Variantie

Voor het berekenen van de variantie wordt er gebruik gemaakt van het factorieel moment [4]. Het factorieel moment stelt dat $E(X^2)$ geschreven kan worden al als $E(X(X-1)) + E(X)$. Dit geldt omdat $E(X(X-1)) + E(X) = E(X^2) - E(X) + E(X) = E(X^2)$. Dit toegepast op de variantie van het Duitse tank probleem geeft

$$\begin{aligned} Var(N|M=m,K=k) = {} & E(N(N-1)|M=m,K=k) \\ & + E(N|M=m,K=k) - E(N|M=m,K=k)^2. \end{aligned}$$

Het uitwerken van $E(N(N-1)|M=m,K=k)$ geeft

$$\begin{aligned} E(N(N-1)|M=m,K=k) &= \sum_{i=m}^{\infty} i(i-1)\cdot P(N=i|M=m,K=k) \\ &= \frac{k-1}{m}\cdot\binom{m}{k}\cdot\sum_{i=m}^{\infty}\frac{i(i-1)}{\binom{i}{k}}, \end{aligned}$$

waarbij geldt dat

$$\begin{aligned} \frac{i(i-1)}{\binom{i}{k}} &= \frac{i(i-1)k!(i-k)!}{i!} \\ &= \frac{k!(i-k)!}{(i-2)!} \\ &= \frac{k(k-1)(k-2)!(i-k)!}{(i-2)!} \\ &= \frac{k(k-1)}{\binom{i-2}{k-2}}. \end{aligned}$$

De identiteit $\sum_{i=m}^{\infty}\binom{i}{k}^{-1} = \frac{m}{k-1}\binom{m}{k}^{-1}$ wordt wederom gebruikt, alleen dit keer met $i-2$ in plaats van i en $k-2$ in plaats van k. Dus

$$\sum_{i=m}^{\infty}\binom{i-2}{k-2}^{-1} = \frac{m-2}{k-3}\binom{m-2}{k-2}^{-1}.$$

De verwachtingswaarde wordt dan

$$\begin{aligned} E(N(N-1)|M=m,K=k) &= \frac{k-1}{m}\cdot\binom{m}{k}\cdot k(k-1)\cdot\frac{m-2}{k-3}\cdot\binom{m-2}{k-2}^{-1} \\ &= \frac{k-1}{m}\cdot k(k-1)\cdot\frac{m-2}{k-3}\cdot\frac{\binom{m}{k}}{\binom{m-2}{k-2}}, \end{aligned}$$

waarbij

$$\begin{aligned} \frac{\binom{m}{k}}{\binom{m-2}{k-2}} &= \frac{\frac{m!}{k!(m-k)!}}{\frac{(m-2)!}{(k-2)!(m-k)!}} \\ &= \frac{m!(k-2)!(m-k)!}{(m-2)!k!(m-k)!} \\ &= \frac{m(m-1)}{k(k-1)}. \end{aligned}$$

Dus

$$\begin{aligned} E(N(N-1)|M=m,K=k) &= \frac{(k-1)k(k-1)(m-2)m(m-1)}{m(k-3)k(k-1)} \\ &= \frac{(k-1)(m-1)(m-2)}{k-3}. \end{aligned}$$

De van $E(N|M=m,K=k)$ is reeds afgeleid, dat is namelijk $(m-1)\frac{k-1}{k-2}$. Hiermee kan ook $E(N|M=m,K=k)^2$ berekend worden:

$$\begin{aligned} E(N|M=m,K=k)^2 &= \left((m-1)\frac{k-1}{k-2}\right)^2 \\ &= (m-1)^2\frac{(k-1)^2}{(k-2)^2} \\ &= (m^2-2m+1)\frac{k^2-2k+1}{k^2-4k+4}. \end{aligned}$$

Wanneer de gevonden verwachtingen ingevuld worden in de functie voor de variantie, volgt

$$\begin{aligned} Var(N|M=m,K=k) &= (m-1)(m-2)\frac{k-1}{k-3}+(m-1)\frac{k-1}{k-2}-(m-1)^2\frac{(k-1)^2}{(k-2)^2} \\ &= (k-1)(m-1)\left(\frac{m-2}{k-3}+\frac{1}{k-2}-\frac{(m-1)(k-1)}{(k-2)^2}\right) \\ &= (k-1)(m-1)\frac{(m-2)(k-2)^2+(k-3)(k-2)-(m-1)(k-1)(k-3)}{(k-3)(k-2)^2}, \end{aligned}$$

waarbij

$$\begin{aligned}
(m-2)(k-2)^2 = (m-2)(k^2-4k+4) &= mk^2-4mk+4m-2k^2+8k-8,\\
(k-3)(k-2) = k^2-2k-3k+6 &= k^2-5k+6,\\
-(m-1)(k-1)(k-3) = -(m-1)(k^2-3k-k+3) &= -mk^2+4mk-3m+k^2-4k+3.
\end{aligned}$$

Dit bij elkaar opgeteld geeft

$$mk^2-4mk+4m-2k^2+8k-8+k^2-5k+6-mk^2+4mk-3m+k^2-4k+3 = 1+m-k,$$

dus de variantie van N in de Bayesiaanse benadering is

$$Var(N|M=m,K=k) = (k-1)(m-1)\frac{1+m-k}{(k-3)(k-2)^2}.$$

Hoofdstuk 5

Vergelijking tussen de benaderingen

5.1 Vergelijking tussen de puntschatters

Voor het vergelijken van de puntschatters wordt gebruik gemaakt van de Gemiddelde Kwadratische Fout (*MSE*). De Gemiddelde Kwadratische Fout is de gemiddelde kwadratische afstand dat de schatter van de ware waarde afligt: $MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2$. De Gemiddelde Kwadratische Fout kan geschreven worden als de som van het kwadraat van de onzuiverheid (*bias*) en de variantie van de schatter, dus

$$MSE[\hat{\theta}] = Bias[\hat{\theta}, \theta]^2 + Var[\hat{\theta}].$$

De onzuiverheid van een puntschatter is gedefinieerd als het verschil tussen de verwachting van de schatter en de ware waarde van de parameter die wordt geschat. Wanneer de verwachting van de schatter gelijk is aan de ware waarde, is de schatter zuiver. Een schatter met een kleinere Gemiddelde Kwadratische Fout ligt gemiddeld dichter bij de ware populatie waarde [6].

Frequentistische benadering

Frequentistische schatters voor onbekende parameters worden geëvalueerd aan de hand van de verdeling van de uitkomsten in de steekproef. Hierbij wordt dus gekeken naar de verdeling van de schatter over alle mogelijke steekproeven. Een belangrijk criterium is dat de schatter zuiver is. Dit houdt in dat de verwachting van de verdeling van de steekproef gelijk is aan de ware (onbekende) populatie verdeling: $E(\hat{n}) = n$. Daarnaast geldt dat een schatter met een lage variantie geprefereerd wordt boven een schatter met een hoge variantie. Daarom wordt in deze scriptie gebruik gemaakt van de zuivere schatter met de kleinste variantie. Het doel van deze schatter is om de waarde van de parameters te vinden die de waarschijnlijkheid van de waargenomen data maximaliseert [6].

In het Duitse tank probleem is de schatter van n gegeven door $\hat{n} = m + \frac{m}{k} - 1$. Met andere woorden, de verwachting van het aantal geproduceerde tanks is gelijk aan de hoogste serienummer in de steekproef plus het gemiddelde gat dat tussen de observaties in de steekproef zit.

Omdat de onzuiverheid voor de Zuivere Schatter met Minimale Variantie logischerwijs gelijk is aan 0, is de Gemiddelde Kwadratische Fout van deze schatter gelijk aan de variantie, ofwel

$$MSE_F(\hat{n}) = Var(\hat{n}) = \frac{(n-k)(n+1)}{k(k+2)}.$$

Bayesiaanse benadering

In tegenstelling tot de frequentistische schatter is de Bayesiaanse *posterior* schatter een onzuivere schatter van n. De onzuiverheid kan namelijk alleen gelijk zijn aan 0 als de verwachting van de *prior* samenvalt met die van de onbekende ware waarde n. De kans hierop is 0. De onzuiverheid van de *posterior* schatter is de verwachte waarde van de schatter min de ware waarde, oftewel $Bias[\hat{n}, n] = E[\hat{n}] - n$. De onzuiverheid wordt lineair groter als de afstand van de verwachting van de *prior* tot de ware verwachting groter wordt.
De verwachting van de gevonden *posterior* schatter minimaliseert de Gemiddelde Kwadratische Fout. Dus $E(N) = (m-1)\frac{(k-1)}{(k-2)}$ is de optimale schatter van n na het meenemen van de data.

Kijkend naar de verwachting van de schatter, waarvoor geldt dat $E(N) = (m-1)\frac{(k-1)}{(k-2)}$, dan is de onzuiverheid

$$Bias[E(N), n] = E[(N)] - n = (m-1)\frac{(k-1)}{(k-2)} - n.$$

In hoofdstuk 4 is de variantie voor de Bayesiaanse schatter afgeleid. Hierbij is

$$Var_B(N) = \frac{(k-1)(m-1)(1+m-k)}{(k-3)(k-2)^2}.$$

De Gemiddelde Kwadratische Fout van de Bayesiaanse schatter is dan

$$MSE_B(N) = \left((m-1)\frac{(k-1)}{(k-2)} - n\right)^2 + \frac{(k-1)(m-1)(1+m-k)}{(k-3)(k-2)^2}.$$

5.2 Vergelijking tussen het betrouwbaarheids- en geloofwaardigheidsinterval

Een andere methode om uitkomsten van schatters te vergelijken is door middel van interval schatters. Interval schatters zijn instrumenten in de statistiek die een bereik van waarden bieden waarin verwacht kan worden dat een parameter zich bevindt. In dit onderdeel worden twee soorten interval schatters behandeld: het betrouwbaarheidsinterval in de frequentistische benadering en het geloofwaardigheidsinterval in de Bayesiaanse benadering.

In de frequentistische benadering wordt de parameter als vast maar onbekend beschouwd. Betrouwbaarheidsintervallen worden voor de steekproef genomen en zijn

gebaseerd op de verdeling van een steekproef van de schatter. Een $(1-\alpha)\times 100\%$ betrouwbaarheidsinterval voor een parameter θ is een interval (l,h) zodat $P(l \leq \theta \leq h) = 1-\alpha$. Dit betekent dat als het experiment herhaald wordt, de verwachting is dat ten minste $(1-\alpha)\times 100\%$ van de intervallen de ware parameter bevatten [6].

In de Bayesiaanse statistiek wordt de onzekerheid over de parameter θ weergegeven door de prior kansverdeling. Het Bayesiaanse geloofwaardigheidsinterval is een interval van parameterwaarden waarvoor de *posterior* kans een vooraf bepaalde waarschijnlijkheid $(1-\alpha)$ heeft. Dit interval wordt meestal gekozen op basis van een zogenaamde Hoogste *Posterior* Dichtheidsfunctie (HPD), die het meest waarschijnlijke bereik van parameterwaarden aangeeft. Dus betekent een $(1-\alpha)\times 100\%$ geloofwaardigheidsinterval dat de kans dat de ware parameterwaarde binnen dat interval ligt gelijk is aan $(1-\alpha)$, gegeven de waargenomen gegevens en de prior informatie [6].

5.3 Resultaten

Er worden vier verschillende benaderingen voor het schatten van het aantal geproduceerde tanks n vergeleken: frequentistische methode, de versimpelde Bayesiaanse methode welke in Hoofdstuk 4 is uitgewerkt, de Bayesiaanse methode met informatieve uniforme prior en de Bayesiaanse methode met een Poisson-verdeling als prior. De evaluatie gebeurt op basis van de verwachtingen, standaarddeviatie (SD), Gemiddelde Kwadratische Fout (MSE) en de betrouwbaarheids- en geloofwaardigheidsintervallen (CI) voor $\alpha = 0.05$. Er wordt vanuit gegaan dat de populatiegrootte n gelijk is aan 1000. In de analyse wordt gekeken naar de resultaten waarbij het aantal gevangen genomen tanks k gelijk is aan 10, 50, 100 en 500. Er worden 10 willekeurige steekproeven gegenereerd per k waarvoor deze gegevens worden berekend. Bij het interpreteren van de uitkomsten wordt gekeken naar het gemiddelde van de 10 steekproeven. Hiervoor is gekozen omdat de resultaten dan minder beïnvloed worden in het geval dat er een uitzonderlijke steekproef gegenereerd wordt. De Poisson *prior* heeft parameter $\lambda = 1000$, wat er voor zorgt dat dit een erg goede schatter is omdat $n = 1000$. In de onderstaande tabellen zijn de resultaten voor de vier verschillende waarden van k te vinden.

Methode	**Verwachting**	**SD**	**MSE**	**CI Onder**	**CI Boven**
Frequent	993.2900	90.87491	8258.2500	936.9662	1049.614
Simple Bayes	1015.7625	127.40583	22261.6482	766.0246	1266.025
Bayes Uniform	1007.9587	104.63445	16150.4245	905.4000	1301.900
Bayes Poisson	996.3768	28.08952	935.5524	945.9000	1054.200

Tabel 5.1: Resultaten voor $k = 10$.

Methode	**Verwachting**	**SD**	**MSE**	**CI Onder**	**CI Boven**
Frequent	1001.1500	19.12459	365.7500	995.8490	1006.451
Simple Bayes	1001.9479	20.36152	505.3869	962.3006	1041.719
Bayes Uniform	1001.9479	20.36151	505.3867	982.5000	1056.400
Bayes Poisson	998.4386	14.06721	256.5794	982.5000	1033.800

Tabel 5.2: Resultaten voor $k = 50$.

Methode	**Verwachting**	**SD**	**MSE**	**CI Onder**	**CI Boven**
Frequent	1001.728	9.398060	88.32353	999.8860	1003.570
Simple Bayes	1001.920	9.705946	139.69790	983.1168	1020.988
Bayes Uniform	1001.920	9.705946	139.69790	992.8000	1027.600
Bayes Poisson	1001.121	8.328443	106.69549	992.8000	1022.800

Tabel 5.3: Resultaten voor $k = 100$.

Methode	**Verwachting**	**SD**	**MSE**	**CI Onder**	**CI Boven**
Frequent	999.998	1.412099	1.994024	999.8742	1000.122
Simple Bayes	1000.004	1.421323	4.028226	997.2164	1002.764
Bayes Uniform	1000.004	1.421323	4.028225	999.0000	1003.000
Bayes Poisson	1000.000	1.414173	3.999877	999.0000	1003.000

Tabel 5.4: Resultaten voor $k = 500$.

Uit de bovenstaande tabellen kunnen verschillende observaties en conclusies worden getrokken.

Gemiddelde Kwadratische Fout (MSE) Voor $k = 10$ is de MSE voor de Frequentistische methode aanzienlijk lager dan voor de Versimpelde en Originele Bayesiaanse methoden, maar hoger dan voor de Bayesiaanse Poisson methode. Naarmate de steekproefgrootte toeneemt, nemen de MSE's van alle methoden af, waarbij de verschillen tussen de methoden kleiner worden. Voor $k = 500$ zijn de MSE's van alle methoden bijna identiek, wat aangeeft dat ze allemaal zeer nauwkeurig zijn bij grote steekproef groottes. Dit is ook logisch omdat bij een steekproef van 500, de helft van de tanks gevangen genomen zijn, en er dus weinig onzekerheid meer is.

Betrouwbaarheids- en Geloofwaardigheidsintervallen (CI) De breedte van het interval geeft de precisie van de schatting aan; een smaller interval betekent een preciezere schatting. Bij $k = 10$ heeft de Bayesiaanse benadering met Poisson prior het smalste interval, wat duidt op een hoge precisie. De Frequentistische methode heeft een breder interval, terwijl de Versimpelde en Originele Bayesiaanse methoden de breedste intervallen hebben, wat wijst op meer onzekerheid. Naarmate k

toeneemt, worden de intervallen voor alle methoden smaller, en bij $k = 500$ zijn de intervallen voor alle methoden bijna identiek. Dit suggereert dat bij grotere steekproeven de verschillende benaderingen dichter bij elkaar komen in termen van precisie.

Bij kleine steekproeven ($k = 10$) biedt de Bayesiaanse Poisson-methode de meest nauwkeurige en precieze schattingen. Dit blijkt uit de laagste Mean Squared Error (MSE) en het smalste betrouwbaarheidsinterval. Naarmate de steekproefgrootte toeneemt, worden de verschillen tussen de methoden kleiner. Bij grote steekproeven ($k = 500$) geven alle methoden vergelijkbare resultaten, met kleine MSE's en smalle intervallen. De frequentistische methode presteert consistent goed, vooral bij grotere steekproefgroottes. Echter, de Bayesiaanse Poisson-methode toont superieure prestaties bij kleinere steekproeven, wat te verklaren valt door de goede prior doordat $\lambda = 1000$. Daarnaast liggen de uitkomsten van de versimpelde Bayesiaanse methode en de Bayesiaanse methode met een uniforme prior waarbij $n_{max} = 1500$, vanaf $k = 50$ erg dicht bij elkaar. Dit betekent dat voor een voldoende grote k de benadering $n_{max} \to \infty$ niet veel slechter is dan een benadering met een informatieve uniforme prior waarbij $n_{max} < \infty$.

5.4 Conclusie

In deze scriptie is een vergelijking gemaakt tussen de freqentistische en Bayesiaanse benaderingen van serienummer analyse in de statistiek aan de hand van het Duitse tank probleem. Hierbij zijn enkele aannames gedaan die de realiteit mogelijk niet volledig weerspiegelen. Er is bijvoorbeeld aangenomen dat de serienummers van de tanks beginnen bij 1 en oplopen naar n in het geval dat er n tanks zijn geproduceerd. In de werkelijkheid bestaat een serienummer vaak uit een vast aantal cijfers of zelfs letters, en is het dus complexer. Daarnaast zou een realistischer model gegenereerd kunnen worden door serienummers van losse onderdelen te combineren. Ook blijkt dat de *prior* kansverdeling veel invloed uitoefent op de kansverdeling van de *posterior* kansfunctie. Het is dus belangrijk om te realiseren dat wanneer je gebruik wilt maken van Bayesiaanse statistiek, er voldoende onderzoek gedaan dient te worden naar een juiste *prior*.

Ondanks deze beperkingen is het Duitse tank probleem een uitermate geschikt onderwerp voor de uitleg en vergelijking van de frequentistische en Bayesiaanse benaderingen binnen de statistiek. Het biedt een duidelijk kader om de verschillen tussen deze benaderingen te illustreren en te begrijpen.

Appendix A

Herhaling regels kansvariabelen

Hieronder volgt voor de volledigheid een kort overzicht van de regels in de statistiek welke als bekend worden beschouwd, maar aangehaald worden in de hoofdstukken.

De voorwaardelijke kans van gebeurtenis B gegeven gebeurtenis A wordt gedefinieerd door

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

De vermenigvuldigingsregel stelt dat

$$P(A \cap B) = P(B) \times P(A|B).$$

De wet van totale kans zegt dat als we te maken hebben met gebeurtenissen $B_1, ..., B_n$ die samen de hele verzameling van mogelijke uitkomsten Ω vormen, dan geldt dat de kans op gebeurtenis A gelijk is aan de som van de kansen van A en elk van de B_j-gebeurtenissen, dus

$$P(A) = \sum_{j=1}^{n} P(B_j \cap A),$$

en als we hier de vermenigvuldigingsregel op toepassen dan krijgen we dat

$$P(A) = \sum_{j} P(B_j) \times P(A|B_j).$$

De verwachtingswaarde van een discrete kansvariabele is gedefinieerd door

$$E[Y] = \sum_{k} y_k P(Y = y_k), \tag{5.1}$$

waar de som wordt genomen over alle mogelijke waarden van de kansvariabele.

De variantie van een discrete kansvariabele wordt gegeven door

$$Var[Y] = E(Y - E[Y])^2 = E[Y^2] - [E[Y]]^2, \tag{5.2}$$

en er geldt dat

$$Var[aY + b] = a^2 Var[Y].$$

Appendix B

R Script

```r
# Laad benodigde libraries
library(ggplot2)
library(stats)

# Functie om de posterior kansfunctie te berekenen voor de originele Bayesiaa
nse benadering
posterior_prob <- function(n, m, k, n_max) {
  likelihood <- choose(m-1, k-1) / choose(n, k)
  prior <- 1 / (n_max + 1)
  likelihood * prior
}

# Functie om de posterior distributie te berekenen voor een gegeven m en k
posterior_distribution <- function(m, k, n_max) {
  n_values <- m:n_max
  posterior <- sapply(n_values, function(n) posterior_prob(n, m, k, n_max))
  posterior <- posterior / sum(posterior)  # Normaliseer de posterior
  return(data.frame(n=n_values, posterior=posterior))
}

# Functie om statistieken te berekenen voor de originele Bayesiaanse benaderi
ng
calculate_statistics <- function(posterior_dist, n_true) {
  expectation <- sum(posterior_dist$n * posterior_dist$posterior)
  variance <- sum((posterior_dist$n - expectation)^2 * posterior_dist$posteri
or)
  sd <- sqrt(variance)
  bias <- expectation - n_true
  mse <- bias^2 + variance
  cumulative <- cumsum(posterior_dist$posterior)
  lower_bound <- posterior_dist$n[which.min(abs(cumulative - 0.025))]
  upper_bound <- posterior_dist$n[which.min(abs(cumulative - 0.975))]
  credible_interval <- c(lower_bound, upper_bound)
  list(expectation=expectation, sd=sd, mse=mse, credible_interval=credible_in
terval)
}

# Functie om de posterior kansfunctie te berekenen voor de Poisson Bayesiaans
e benadering
posterior_prob_poisson <- function(n, m, k, lambda) {
  likelihood <- choose(m-1, k-1) / choose(n, k)
  prior <- dpois(n, lambda)
  likelihood * prior
}

# Functie om de posterior distributie te berekenen voor de Poisson Bayesiaans
e benadering
posterior_distribution_poisson <- function(m, k, lambda, n_max) {
  n_values <- m:n_max
  posterior <- sapply(n_values, function(n) posterior_prob_poisson(n, m, k, l
```

```
ambda))
  posterior <- posterior / sum(posterior)  # Normaliseer de posterior
  return(data.frame(n=n_values, posterior=posterior))
}

# Functie om de credible interval te berekenen
credible_interval <- function(values, probs=c(0.025, 0.975)) {
  quantile(values, probs)
}

# Functie om de confidence interval te berekenen
confidence_interval <- function(mean, sd, n, alpha=0.05) {
  error_margin <- qnorm(1 - alpha/2) * sd / sqrt(n)
  c(mean - error_margin, mean + error_margin)
}

# Parameters
n_max <- 1500
n <- 1000
k_values <- c(10, 50, 100, 500)
num_samples <- 10
lambda <- 1000  # Parameter voor de Poisson-verdeling

# Functie om 10 willekeurige steekproeven te genereren en de resultaten te be
rekenen
analyze_samples <- function(k, n_max, n_true, num_samples) {
  results <- replicate(num_samples, {
    # Generate een willeukerige steekproef
    random_sample <- sample(1:1000, k, replace=FALSE)
    m <- max(random_sample)

    # Bereken statistieken voor de frequentistische benadering
    frequentist_exp <- m + m / k - 1
    frequentist_var <- (n - k) * (n + 1) / (k * (k + 2))
    frequentist_sd <- sqrt(frequentist_var)
    frequentist_mse <- (n - k) * ( n + 1) / (k * (k + 2))
    frequentist_ci <- confidence_interval(frequentist_exp, frequentist_sd, k)

    # Bereken statistieken voor de versimpelde Bayesiaanse benadering
    simple_bayes_expectation <- (m - 1) * (k - 1) / (k - 2)
    simple_bayes_variance <- ((k - 1) * (m - 1) * (1 + m - k)) / ((k - 3) * (
k - 2)^2)
    simple_bayes_sd <- sqrt(simple_bayes_variance)
    simple_bayes_mse <- ((m - 1) * (k - 1) / (k - 2) - n)^2 + (k - 1) * (m -
1) * (1 + m - k) / ((k - 3) * (k - 2)^2)
    simple_bayes_ci <- credible_interval(rnorm(10000, simple_bayes_expectatio
n, simple_bayes_sd))
```

```r
# Bereken de posterior distributie en statistieken voor de uniforme Bayesiaan
se benadering
    posterior_dist <- posterior_distribution(m, k, n_max)
    bayes_stats <- calculate_statistics(posterior_dist, n_true)

    # Bereken de posterior distributie en statistieken voor de Poisson Bayesi
aanse benadering
    posterior_dist_poisson <- posterior_distribution_poisson(m, k, lambda, n_
max)
    bayes_poisson_stats <- calculate_statistics(posterior_dist_poisson, n_tru
e)

    list(
      m = m,
      simple_bayes = list(expectation = simple_bayes_expectation, sd = simple
_bayes_sd, mse = simple_bayes_mse, ci = simple_bayes_ci),
      frequentist = list(expectation = frequentist_exp, sd = frequentist_sd,
mse = frequentist_mse, ci = frequentist_ci),
      bayesian = bayes_stats,
      bayesian_poisson = bayes_poisson_stats
    )
  }, simplify=FALSE)
  return(results)
}

# Berekeningen voor verschillende k waarden
all_results <- lapply(k_values, function(k) {
  list(k = k, results = analyze_samples(k, n_max, n, num_samples))
})

# Functie om de resultaten te verzamelen en te tonen in een tabel
generate_result_table <- function(results) {
  result_table <- data.frame(
    Method = character(),
    Expectation = numeric(),
    SD = numeric(),
    MSE = numeric(),
    CI_Lower = numeric(),
    CI_Upper = numeric()
  )

  frequentist_means <- sapply(results$results, function(res) res$frequentist$
expectation)
  frequentist_sds <- sapply(results$results, function(res) res$frequentist$sd
)
  frequentist_mses <- sapply(results$results, function(res) res$frequentist$m
se)
  frequentist_cis <- t(sapply(results$results, function(res) res$frequentist$
ci))
```

```r
  avg_frequentist_mean <- mean(frequentist_means)
  avg_frequentist_sd <- mean(frequentist_sds)
  avg_frequentist_mse <- mean(frequentist_mses)
  avg_frequentist_ci <- apply(frequentist_cis, 2, mean)

  result_table <- rbind(result_table, data.frame(
    Method = "Frequentist",
    Expectation = avg_frequentist_mean,
    SD = avg_frequentist_sd,
    MSE = avg_frequentist_mse,
    CI_Lower = avg_frequentist_ci[1],
    CI_Upper = avg_frequentist_ci[2]
  ))


  simple_bayes_means <- sapply(results$results, function(res) res$simple_baye
s$expectation)
  simple_bayes_sds <- sapply(results$results, function(res) res$simple_bayes$
sd)
  simple_bayes_mses <- sapply(results$results, function(res) res$simple_bayes
$mse)
  simple_bayes_cis <- t(sapply(results$results, function(res) res$simple_baye
s$ci))

  avg_simple_bayes_mean <- mean(simple_bayes_means)
  avg_simple_bayes_sd <- mean(simple_bayes_sds)
  avg_simple_bayes_mse <- mean(simple_bayes_mses)
  avg_simple_bayes_ci <- apply(simple_bayes_cis, 2, mean)

  result_table <- rbind(result_table, data.frame(
    Method = "Simplified Bayesian",
    Expectation = avg_simple_bayes_mean,
    SD = avg_simple_bayes_sd,
    MSE = avg_simple_bayes_mse,
    CI_Lower = avg_simple_bayes_ci[1],
    CI_Upper = avg_simple_bayes_ci[2]
  ))

  bayes_means <- sapply(results$results, function(res) res$bayesian$expectati
on)
  bayes_sds <- sapply(results$results, function(res) res$bayesian$sd)
  bayes_mses <- sapply(results$results, function(res) res$bayesian$mse)
  bayes_cis <- t(sapply(results$results, function(res) res$bayesian$credible_
interval))

  avg_bayes_mean <- mean(bayes_means)
  avg_bayes_sd <- mean(bayes_sds)
  avg_bayes_mse <- mean(bayes_mses)
```

```r
  avg_bayes_ci <- apply(bayes_cis, 2, mean)

  result_table <- rbind(result_table, data.frame(
    Method = "Bayesian",
    Expectation = avg_bayes_mean,
    SD = avg_bayes_sd,
    MSE = avg_bayes_mse,
    CI_Lower = avg_bayes_ci[1],
    CI_Upper = avg_bayes_ci[2]
  ))

  bayes_poisson_means <- sapply(results$results, function(res) res$bayesian_p
oisson$expectation)
  bayes_poisson_sds <- sapply(results$results, function(res) res$bayesian_poi
sson$sd)
  bayes_poisson_mses <- sapply(results$results, function(res) res$bayesian_po
isson$mse)
  bayes_poisson_cis <- t(sapply(results$results, function(res) res$bayesian_p
oisson$credible_interval))

  avg_bayes_poisson_mean <- mean(bayes_poisson_means)
  avg_bayes_poisson_sd <- mean(bayes_poisson_sds)
  avg_bayes_poisson_mse <- mean(bayes_poisson_mses)
  avg_bayes_poisson_ci <- apply(bayes_poisson_cis, 2, mean)

  result_table <- rbind(result_table, data.frame(
    Method = "Bayesian Poisson",
    Expectation = avg_bayes_poisson_mean,
    SD = avg_bayes_poisson_sd,
    MSE = avg_bayes_poisson_mse,
    CI_Lower = avg_bayes_poisson_ci[1],
    CI_Upper = avg_bayes_poisson_ci[2]
  ))

  return(result_table)
}

# Genereren van tabellen voor elke waarde van k
for (result in all_results) {
  cat(sprintf("\nResultaten voor k=%d\n", result$k))
  print(generate_result_table(result))
}

##
## Resultaten voor k=10
##                 Method  Expectation       SD        MSE CI_Lower CI_Upper
##             Frequentist  993.2900   90.87491  8258.2500 936.9662 1049.614
##    Simplified Bayesian  1015.7625 127.40583 22261.6482 766.0246 1266.025
##                Bayesian  1007.9587 104.63445 16150.4245 905.4000 1301.900
##       Bayesian Poisson  996.3768   28.08952   935.5524 945.9000 1054.200
```

```
##
## Resultaten voor k=50
##                   Method Expectation       SD      MSE CI_Lower CI_Upper
##              Frequentist   1001.1500 19.12459 365.7500 995.8490 1006.451
##      Simplified Bayesian   1001.9479 20.36152 505.3869 962.3006 1041.719
##                  Bayesian   1001.9479 20.36151 505.3867 982.5000 1056.400
##         Bayesian Poisson    998.4386 14.06721 256.5794 982.5000 1033.800
##
## Resultaten voor k=100
##                   Method Expectation       SD       MSE CI_Lower CI_Upper
##              Frequentist    1001.728 9.398060  88.32353 999.8860 1003.570
##      Simplified Bayesian    1001.920 9.705946 139.69790 983.1168 1020.988
##                  Bayesian    1001.920 9.705946 139.69790 992.8000 1027.600
##         Bayesian Poisson    1001.121 8.328443 106.69549 992.8000 1022.800
##
## Resultaten voor k=500
##                   Method Expectation       SD      MSE CI_Lower CI_Upper
##              Frequentist     999.998 1.412099 1.994024 999.8742 1000.122
##      Simplified Bayesian    1000.004 1.421323 4.028226 997.2164 1002.764
##                  Bayesian    1000.004 1.421323 4.028225 999.0000 1003.000
##         Bayesian Poisson    1000.000 1.414173 3.999877 999.0000 1003.000
```

Bibliografie

[1] Hirotugu Akaike. "The Interpretation of Improper Prior Distributions as Limits of Data Dependent Proper Prior Distributions". In: *Journal of the Royal Statistical Society. Series B (Methodological), Vol. 42, No. 1* (1980), p. 46–52. URL: https://www.jstor.org/stable/2984737.

[2] David Blackwell. "Conditional Expectation and Unbiased Sequential Estimation". In: *Ann. Math. Statist. 18(1)* (Maart 1947), p. 105–110. DOI: 10.1214/aoms/1177730497.

[3] Prof. Dr. J. Boersma. *Voortgezette Functietheorie*. 1978-1984. URL: %5Curl%7Bhttps://casa.win.tue.nl/education/AntWiskDict/_3/a.%20Analyse%20en%20toepassingen%20van%20de%20Analyse/13.ManuscriptVoortgezetteFunctietheor
1984).pdf%7D.

[4] D. S. Broca. "Mean and Variance through Factorial Moments". In: *Teaching Statistics, Volume 27, Issue 2* (2005), p. 55–57. DOI: 10.1111/j.1467-9639.2005.00209.x.

[5] R.Ruggles; H. Brodie. "An Empirical Approach to Economic Intelligence in World War II". In: *Journal of the American Statistical Association, Vol. 42, No. 237* (1947), p. 72–91. DOI: 10.2307/2280189.

[6] William M. Bolstad; James M. Curran. *Introduction to Bayesian Statistics*. Wiley, 2017. DOI: 10.1002/9781118593165.

[7] Leo A. Goodman. "Serial Number Analysis". In: *Journal of the American Statistical Association Vol. 47, No. 260* (Dec 1952), p. 622–634. URL: https://www.jstor.org/stable/2280780.

[8] Höhl; Held. "Bayesian Estimation of the Size of a Population". In: *Sonderforschungsbereich 386, Paper 499* (2006). DOI: 10.5282/ubm/epub.2094.

[9] Micheal Chimaobi Kalu. "How the Allies Guessed the Number of German Tanks Using Serial Numbers". In: *War History Online* (Jan 26, 2019). URL: https://www.warhistoryonline.com/instant-articles/the-german-tank-problem.html.

[10] Kazuhiko Aomoto; Michitake Kita. *Theory of Hypergeometric Functions*. Springer, 2011. Hfdstk. 1, p. 1. DOI: 10.1007/978-4-431-53938-4.

[11] *Lecture notes on Mathematical Statistics WISB263*. Utrecht University, 2023.

[12] George Clark; Alex Gonye; Steven J Miller. *Lessons from the German Tank Problem*. 2021. DOI: 10.48550/arXiv.2101.08162.

[13] Anthony O'Hagan. *Handbook of Probability: Theory and Applications*. Sage Publications, Inc., 2008. Hfdstk. 6, p. 85. DOI: 10.4135/9781452226620.

[14] Anthony O'Hagan. *Kendall's advanced theory of statistics 2B - Bayesian inferencex*. John Wiley Sons Ltd, 1999. DOI: 10.1111/j.1467-985X.2004.00347_15.x.

[15] F. Bartolucci; L. Scrucca. *International Encyclopedia of Education (Third edition)*. Elsevier, 2010, p. 366–373. DOI: 10.1016/B978-0-08-044894-7.01376-2.

[16] Cory M. Simon. "A Bayesian Treatment of the German Tank Problem". In: *The Mathematical Intelligencer* (2023). DOI: 10.1007/s00283-023-10274-6.

[17] Gordon A. Fox; Simoneta Negrete-Yankelevich; Vinicio J. Sosa. *Ecological Statistics: Contemporary theory and application*. Oxford University Press, 2015. Hfdstk. 1, p. 7. DOI: 10.1093/acprof:oso/9780199672547.001.0001.