STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING
LEIDING: PROF. DR D. VAN DANTZIG
ADVISEUR VOOR STATISTISCHE CONSULTATIE: PROF. DR J. HEMELRIJK

Rapport S 249

De invloed van een prioriteitsregeling op de gemiddelde wachttijd

door
F.W. Steutel

april 1959

De invloed van een prioriteitsregeling op de gemiddelde wachttijd.

In veel situaties, waarbij wachttijden optreden, is het van belang, dat de gemiddelde wachttijd zo klein mogelijk wordt gehouden.

In het volgende zullen wij de terminologie gebruiken, die betrekking heeft op klanten, die aankomen voor een loket; de theorie is echter op tal van analoge situaties van toepassing. Men denkt bijv. aan auto's, die voor reparatie in een garage komen, schepen, die aankomen in een haven e.d.

Wij beschouwen de volgende situatie: klanten, die wij in de volgorde van hun aankomst genummerd denken (1, 2, 3, ...), komen aan voor een loket, zodanig dat de tijdsintervallen $\underline{y}_n$ tussen de n^e en de $(n+1)^e$ aankomst de (aankomstintervallen) onderling onafhankelijke grootheden zijn, alle met dezelfde exponentiële verdeling $A(y)$ met gemiddelde $\mathcal{E}\underline{y} = \frac{1}{\lambda}$, dus

$$A(y) \begin{cases} 1 - e^{-\lambda y} & \text{als } y \geq 0 \\ 0 & \text{als } y < 0 \end{cases}$$

λ geeft tevens het gemiddelde aantal aankomsten per tijdseenheid aan.

De tijd, nodig om de n^e klant aan het loket te behandelen, noemen wij zijn bedieningstijd, die wij aangeven met $\underline{s}_n$. Wij veronderstellen, dat de bedieningstijden $\underline{s}_n$ onderling onafhankelijke grootheden zijn, die alle dezelfde verdeling $B(s)$ bezitten met gemiddelde $\mathcal{E}\underline{s} = \mu$. Bovendien veronderstellen wij, dat de bedieningstijden en de aankomstintervallen onderling onafhankelijk zijn.

De verhouding $\frac{\mathcal{E}\underline{s}}{\mathcal{E}\underline{y}} = \lambda\mu$ noemen wij de bezettingsgraad, die wij aangeven met ρ. Als $\rho < 1$ is, ontstaat na een zekere aanlooptijd, gerekend vanaf de aanvang van het proces, een evenwichtssituatie (die niet afhangt van de situatie, die de eerste klant aan het loket aantrof), waarin de wachttijd, die op grond van de bovengenoemde veronderstellingen een kansverdeling bezit, voor iedere klant dezelfde verdeling heeft, zodat ook de verwachting van de wachttijd voor iedere klant hetzelfde is. Deze aanlooptijd is wiskundig gezien oneindig, doch in praktische gevallen zal deze evenwichtssituatie na een eindige tijd reeds bij voldoende goede

benadering bereikt zijn. Wij zullen verder alleen deze evenwichtssituatie beschouwen. De gemiddelde wachttijd van een klant hangt dan af van de factoren

1) de verdelingsfunctie B(s) van de bedieningstijd;
2) de volgorde, waarin de klanten behandeld worden.

De factor 1) zal men in het algemeen niet kunnen wijzigen, daar bij een gegeven soort klanten en een gegeven bedieningsmechanisme B(s) vast ligt. Men kan de gemiddelde wachttijd echter wel beïnvloeden door de factor 2) te wijzigen, bijvoorbeeld door bepaalde klanten voorrang te verlenen, in plaats van, zoals dikwijls wordt gedaan, alle klanten in volgorde van aankomst te behandelen. Er blijkt altijd een vermindering van de gemiddelde wachttijd op te treden, wanneer men bij de behandeling van twee klanten, waarvan de eerst aangekomene de langste bedieningstijd heeft, de volgorde verwisselt.

Wij beschouwen eerst het geval, dat de klanten verdeeld worden in twee prioriteitsklassen, waarbij klanten van de eerste prioriteit voorrang hebben voor klanten van de tweede prioriteit en klanten van dezelfde prioriteit in volgorde van aankomst worden behandeld.

Wij kunnen hierbij twee gevallen onderscheiden:

I De klanten zijn op natuurlijke wijze in twee groepen verdeeld, waarbij de bedieningstijd $\underline{s}_1$ van klanten van de eerste soort de verdeling $B_1(s)$ heeft en de bedieningstijd $\underline{s}_2$ van klanten van de tweede soort de verdeling $B_2(s)$. Wij veronderstellen, dat de klanten van de eerste en tweede soort onafhankelijk van elkaar aankomen en dat de aankomstintervallen $\underline{y}_{(1)}$ resp. $\underline{y}_{(2)}$ voor beide soorten klanten exponentieel verdeeld zijn met gemiddelden $\frac{1}{\lambda_1}$ resp. $\frac{1}{\lambda_2}$, dus

$$A_1(y) = 1 - e^{-\lambda_1 y}$$
$$A_2(y) = 1 - e^{-\lambda_2 y}.$$

Men kan dit ook als volgt interpreteren: voor iedere klant wordt onafhankelijk met kansen $\frac{\lambda_1}{\lambda_1+\lambda_2}$ resp. $\frac{\lambda_2}{\lambda_1+\lambda_2}$ geloot, of hij van de eerste resp. tweede soort is.

II De klanten zijn niet in twee groepen te onderscheiden.

Ad I

De veronderstellingen onder I houden in, dat het aankomstinterval voor alle klanten (d.w.z. het tijdsinterval tussen twee opeenvolgende aankomsten van willekeurige klanten, afgezien van de groep waartoe zij behoren) weer exponentieel verdeeld is, dus

$$A(y) = 1 - e^{-\lambda y},$$

waarin nu $\lambda = \lambda_1 + \lambda_2$ is.

De verdelingsfunctie van de bedieningstijd voor een willekeurige klant kan men nu schrijven als

$$B(s) = \frac{\lambda_1}{\lambda} B_1(s) + \frac{\lambda_2}{\lambda} B_2(s),$$

zodat

$$\mathcal{E}\,\underline{s} = \frac{\lambda_1}{\lambda}\,\mathcal{E}\,\underline{s}_1 + \frac{\lambda_2}{\lambda}\,\mathcal{E}\,\underline{s}_2 \qquad \text{of} \qquad \mu = \frac{\lambda_1}{\lambda}\,\mu_1 + \frac{\lambda_2}{\lambda}\,\mu_2$$

Om de verwachting van de wachttijd te berekenen, maken wij gebruik van de formules van A. Cobham[1]. Voor het geval van twee prioriteiten luiden deze

$$(1) \qquad \begin{cases} \mathcal{E}\,\underline{w}_1 = \dfrac{\lambda\,\mathcal{E}\,\underline{s}^2}{2(1-\lambda_1\mu_1)} \\ \mathcal{E}\,\underline{w}_2 = \dfrac{\lambda\,\mathcal{E}\,\underline{s}^2}{2(1-\lambda_1\mu_1)(1-\rho)} \end{cases},$$

waarin $\mathcal{E}\,\underline{w}_1$ de gemiddelde wachttijd van de eerste prioriteit en $\mathcal{E}\,\underline{w}_2$ de gemiddelde wachttijd van de tweede prioriteit voorstelt. Wij zullen nu nagaan, of wij de gemiddelde wachttijd voor alle klanten kunnen verminderen door aan de klanten van één van beide groepen voorrang te verlenen.

Men ziet uit (1), dat de verhouding van de gemiddelde wachttijden voor beide prioriteitsklassen alleen van de bezettingsgraad ρ afhangt, nl.

---

1) A. Cobham, Priority assignment in waiting line problems, J.Op. Res. Soc. of America, 2, 70-76 (1954).
A. Cobham, Priority assignment - a correction, J.Op.Res. Soc. of America, 3, 547-547 (1955).

$$(2) \qquad \frac{\mathcal{E}\,\underline{w}_1}{\mathcal{E}\,\underline{w}_2} = 1 - \rho$$

Wij willen hier speciaal de gemiddelde wachttijd $\mathcal{E}\,\underline{w}^*$ van een willekeurige klant in het geval van twee prioriteitsklassen vergelijken met de gemiddelde wachttijd $\mathcal{E}\,\underline{w}$ van een willekeurige klant bij bediening in volgorde van aankomst. Uit (1) volgt, dat de gemiddelde wachttijd van een willekeurige klant bij invoering van twee prioriteiten gegeven wordt door

$$(3) \qquad \mathcal{E}\,\underline{w}^* = \frac{\lambda_1}{\lambda}\,\mathcal{E}\,\underline{w}_1 + \frac{\lambda_2}{\lambda}\,\mathcal{E}\,\underline{w}_2 = \frac{\lambda\,\mathcal{E}\,\underline{s}^2}{2(1-\rho)}\,\frac{1-\lambda_1\mu}{1-\lambda_1\mu_1} \,.$$

De gemiddelde wachttijd bij bediening in volgorde van aankomst is

$$(4) \qquad \mathcal{E}\,\underline{w} = \frac{\lambda\,\mathcal{E}\,\underline{s}^2}{2(1-\rho)} \,,$$

zodat volgens (3)

$$(5) \qquad \frac{\mathcal{E}\,\underline{w}^*}{\mathcal{E}\,\underline{w}} = \frac{1-\lambda_1\mu}{1-\lambda_1\mu_1} \,.$$

Uit (5) blijkt, dat men om een vermindering van de gemiddelde wachttijd te bereiken, voorrang moet verlenen aan de klanten uit de groep met de kleinste gemiddelde bedieningstijd, daar als $\mu_1 < \mu_2$ is, tevens $\mu_1 < \mu$ is, zodat dan $\frac{1-\lambda_1\mu}{1-\lambda_1\mu_1}$ is.

Uit (5) volgt verder, dat bij een hogere bezettingsgraad een grotere vermindering van de wachttijd wordt bereikt. Immers, houdt men μ_1, μ_2 en de verhouding $\frac{\lambda_1}{\lambda_2}$ constant, zodat dus μ constant is en laat men het aantal aankomsten per tijdseenheid λ toenemen, zodat ook λ_1 toeneemt, dan neemt (als $\mu_1 < \mu$) het quotient $\frac{1-\lambda_1\mu}{1-\lambda_1\mu_1}$ af.

Ad II

Het geval, waarbij geen splitsing van de klanten in twee groepen bestaat, kan men tot geval I herleiden, door nu, zoals

door H.J. Prins is gedaan (in een, voorzover mij bekend, niet gepubliceerd rapport) voor het geval van een exponentiële verdeling voor de bedieningstijd, de klanten te splitsen in die met een bedieningstijd $\leq c\mu$ en die met een bedieningstijd $> c\mu$. Het gaat er dan om, de constante c zo te bepalen, dat de gemiddelde wachttijd minimaal wordt. Hierbij is ondersteld, dat de bedieningstijd van een klant voldoende nauwkeurig geschat kan worden, hetgeen in praktische situaties vaak het geval zal zijn.

Door deze splitsing ontstaat een situatie, waarbij aan de voorwaarden onder I voldaan is, zodat wij weer gebruik kunnen maken van de formules (1) t/m (5) waarbij

$$\lambda_1 = \lambda B(c\mu)_1 \quad \lambda_2 = \lambda(1 - B(c\mu))_1 \quad \mu_1 = \int_0^c \frac{s\,dB(s)}{B(c\mu)} \quad \text{en}$$

$$\mu_2 = \int_{c\mu}^{\infty} \frac{s\,dB(s)}{1-B(c\mu)} \quad \text{genomen moeten worden.}$$

Wij beschouwen speciaal formule (5), deze wordt nu

$$\frac{\mathcal{E}_{W^*}}{\mathcal{E}_W} = \frac{1-\rho B(c\mu)}{1-\lambda \int_0^c s\,dB(s)} \tag{5a}$$

Voorbeeld:

Bij een exponentiële verdeling voor de bedieningstijd met gemiddelde μ, dus $B(s) = 1 - e^{-\frac{s}{\mu}}$, vindt men

$$\frac{\mathcal{E}_{W^*}}{\mathcal{E}_W} = \frac{1-\rho(1-e^{-c})}{1-\rho(1-e^{-c}-ce^{-c})} \tag{5b}$$

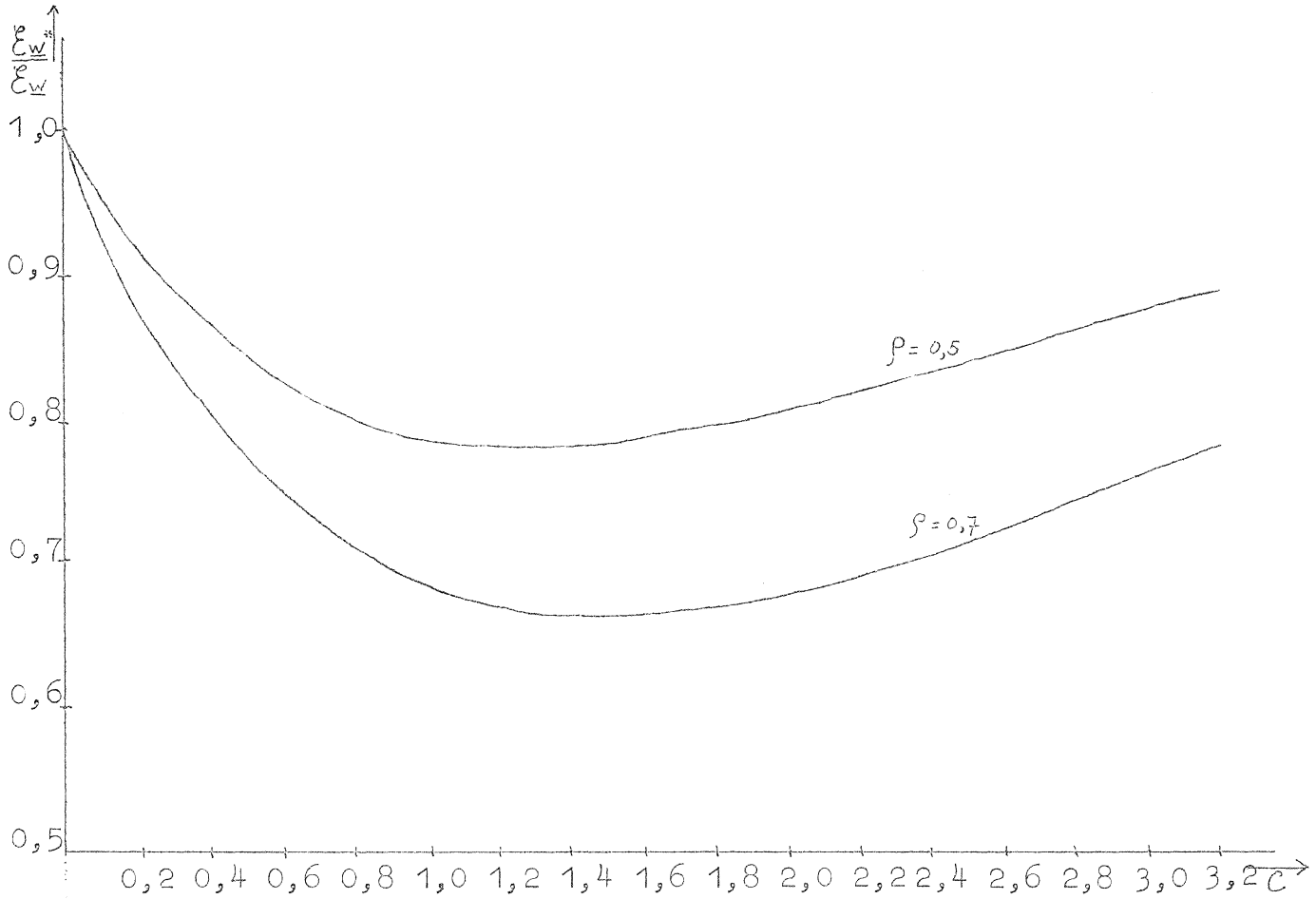
In figuur 1 is deze verhouding uitgezet als functie van c voor $\rho = 0{,}5$ en $\rho = 0{,}7$.

Als wij onderstellen, dat de functie B(s) differentieerbaar[1] is, kunnen wij door differentiatie van (5a) naar c het minimum van $\frac{\mathcal{E}_{W^*}}{\mathcal{E}_W}$ vinden. Dit minimum blijkt altijd te bestaan en eenduidig bepaald te zijn. Als wij de waarde van c, waarvoor $\frac{\mathcal{E}_{W^*}}{\mathcal{E}_W}$ minimaal

---

1) Als B(s) niet differentieerbaar is, kan men het minimum van $\frac{\mathcal{E}_{W^*}}{\mathcal{E}_W}$ uit de grafiek van (5a) bepalen.

Figuur 1



$$\frac{\mathcal{E}_{\underline{w}}^{*}}{\mathcal{E}_{\underline{w}}} = \frac{1-\rho(1-e^{-c})}{1-\rho(1-e^{-c}-ce^{-c})}$$ als functie van c voor $\rho = 0,5$ en $\rho = 0,7$.

wordt, aangeven met c^*, dan vinden wij hieruit op zeer eenvoudige wijze de waarde van dit minimum. Bij berekening blijkt namelijk, dat

$$(6) \qquad \frac{\varphi_{\underline{W}^*}}{\varphi_{\underline{W}}}\ \text{min.} = \frac{1}{c^*}$$

is.

Men bepaalt c^* uit de vergelijking

$$(7) \qquad c^* = 1 + \lambda \int_0^{c^*\mu} B(s)ds = 1 + \rho \int_0^{c^*} B(\mu s)ds$$

zodat steeds $c^* > 1$ is. Om het optimale resultaat te bereiken, moet men de splitsing van de bedieningstijden dus <u>boven het gemiddelde</u> aanbrengen.

Men ziet uit formule (7) (bijv. uit het feit, dat $\frac{\partial c^*}{\partial \rho} > 0$ is) dat c^* toeneemt met toenemende ρ zodat dus $\frac{\varphi_{\underline{W}^*}}{\varphi_{\underline{W}}} = \frac{1}{c^*}$ afneemt met toenemende ρ. <u>Hoe hoger dus de bezettingsgraad is, des te groter is de optimale reductie van de gemiddelde wachttijd</u>.

<u>Voorbeelden</u>:

Als $B(s) = 1 - e^{-\frac{s}{\mu}}$ is, dan vindt men uit (7) de vergelijking

$$(8) \qquad c^* = 1 + \frac{\rho}{1-\rho}\ e^{-c^*},$$

waaruit c^* door iteratie gemakkelijk opgelost kan worden.

Als $B(s)$ een homogene verdeling is op het interval $(\mu-h, \mu+h)$, waarbij $h \leqslant \mu$ is, dus

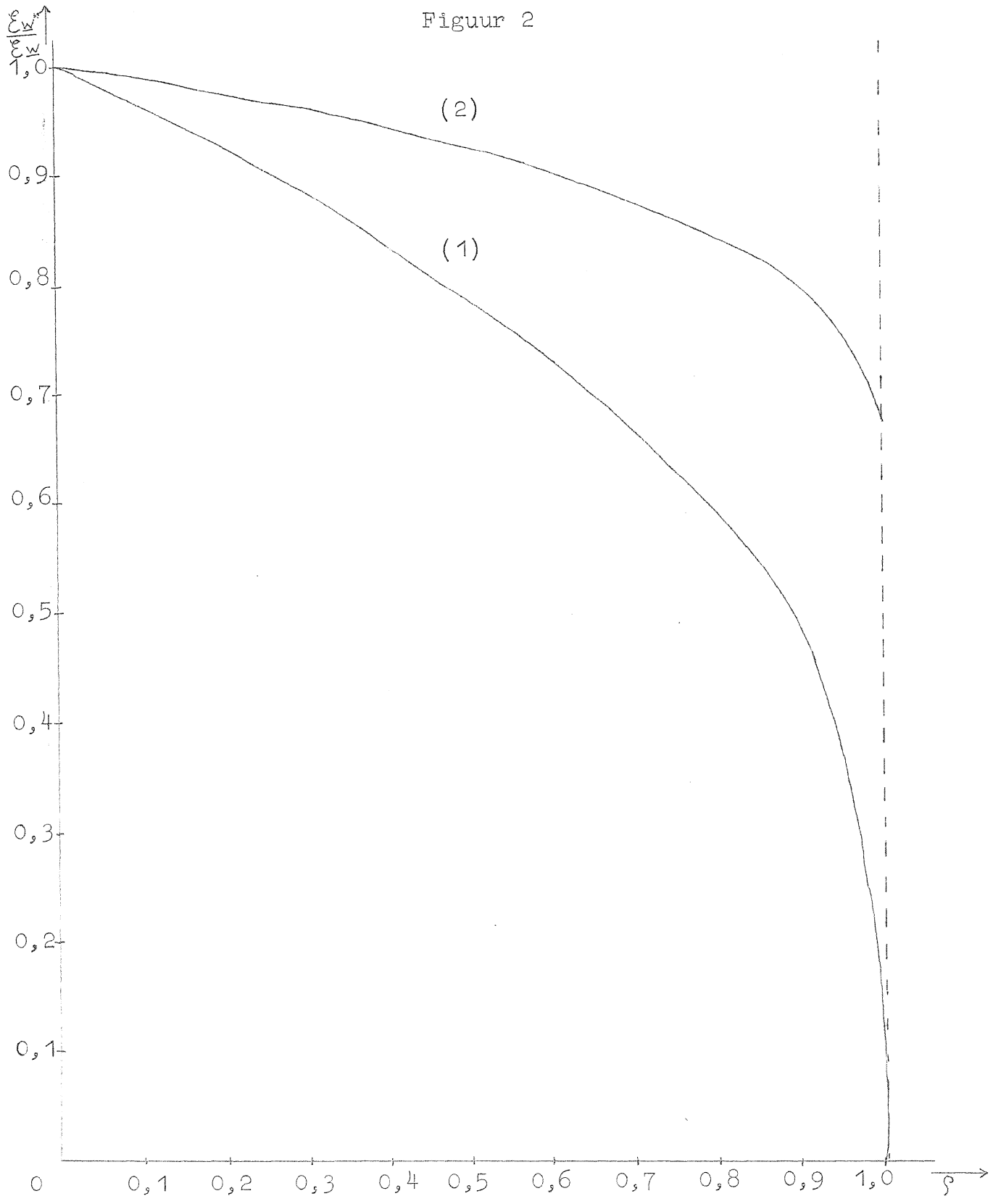
$$B(s) = \begin{cases} 0 & \text{als } s \leqslant \mu - h \\ \frac{s-\mu+h}{2h} & \text{als } \mu - h < s \leqslant \mu + h \\ 1 & \text{als } s > \mu + h\ , \end{cases}$$

dan wordt (7)

$$(9) \qquad c^* = \frac{\rho\mu + h(2 - \rho - 2\sqrt{1-\rho})}{\rho\mu}$$

In figuur (2) is $\frac{\varphi_{\underline{W}^*}}{\varphi_{\underline{W}}} = \frac{1}{c^*}$ getekend voor het geval dat $B(s) = 1 - e^{-\frac{s}{\mu}}$ en voor het geval, dat $B(s)$ een homogene verdeling is

Figuur 2



(1) $\left(\frac{\varepsilon_{W^*}}{\varepsilon_W}\right)_{min} = \frac{1}{c^*}$, met $c^* = 1 + \frac{\rho}{1-\rho} e^{-c^*}$ als functie van ρ.

(2) $\left(\frac{\varepsilon_{W^*}}{\varepsilon_W}\right)_{min} = \frac{\rho}{\rho + 2 - 2\sqrt{1-\rho}}$ als functie van ρ.

op het interval $(\frac{1}{2}, \frac{3}{2})$ als functie van ρ.

Uit figuur 2 blijkt, dat in het geval van een exponentiële verdeling van de bedieningstijd het quotiënt $\frac{\mathcal{E}_{\underline{W}^*}}{\mathcal{E}_{\underline{W}}}$ voor $\rho \to 1$ naar nul gaat en dat in het geval van de homogene verdeling dit quotiënt groter dan $\frac{2}{3}$ blijft. Algemeen[1] geldt, dat $\frac{\mathcal{E}_{\underline{W}^*}}{\mathcal{E}_{\underline{W}}}$ naar $\frac{\mu}{M}$ gaat voor ρ naar 1 gaande, als de bedieningstijd waarden $\leq M$ kan aannemen. Als de bedieningstijd willekeurig grote waarden kan aannemen, gaat $\frac{\mathcal{E}_{\underline{W}^*}}{\mathcal{E}_{\underline{W}}}$ dus naar nul voor ρ naar 1 gaande.

Oneindig veel prioriteiten

Men kan de klanten op analoge wijze in meer dan twee groepen verdelen. Als men binnen een bepaalde splitsing een verdere splitsing aanbrengt, heeft dat een verdere vermindering van de wachttijd tot gevolg. De grootst mogelijke vermindering van de gemiddelde wachttijd bereikt men, door steeds de klant met de kleinste bedieningstijd het eerst te behandelen. Dit geval is beschouwd door T.E. Phipps[2]. Om deze methode te kunnen toepassen, zou men dus de bedieningstijd van iedere klant redelijk moeten kunnen schatten, hetgeen in de praktijk niet altijd mogelijk zal zijn.

Men vindt in dit geval voor de verhouding van de gemiddelde wachttijd $\mathcal{E}_{\underline{W}^{**}}$ hg bediening volgens bovenstaande methode tot de gemiddelde wachttijd bij bediening in volgorde van aankomst

$$\frac{\mathcal{E}_{\underline{W}^{**}}}{\mathcal{E}_{\underline{W}}} = \frac{1}{1-\rho} \int_0^\infty \frac{dB(s)}{\left\{1-\lambda \int_0^s x\,dB(x)\right\}^2} \qquad (10)$$

In figuur 3 is deze verhouding getekend als functie van voor een exponentiële verdeling van de bedieningstijd en voor een homogeen verdeelde bedieningstijd op het interval $(\frac{1}{2}, \frac{3}{2})$.

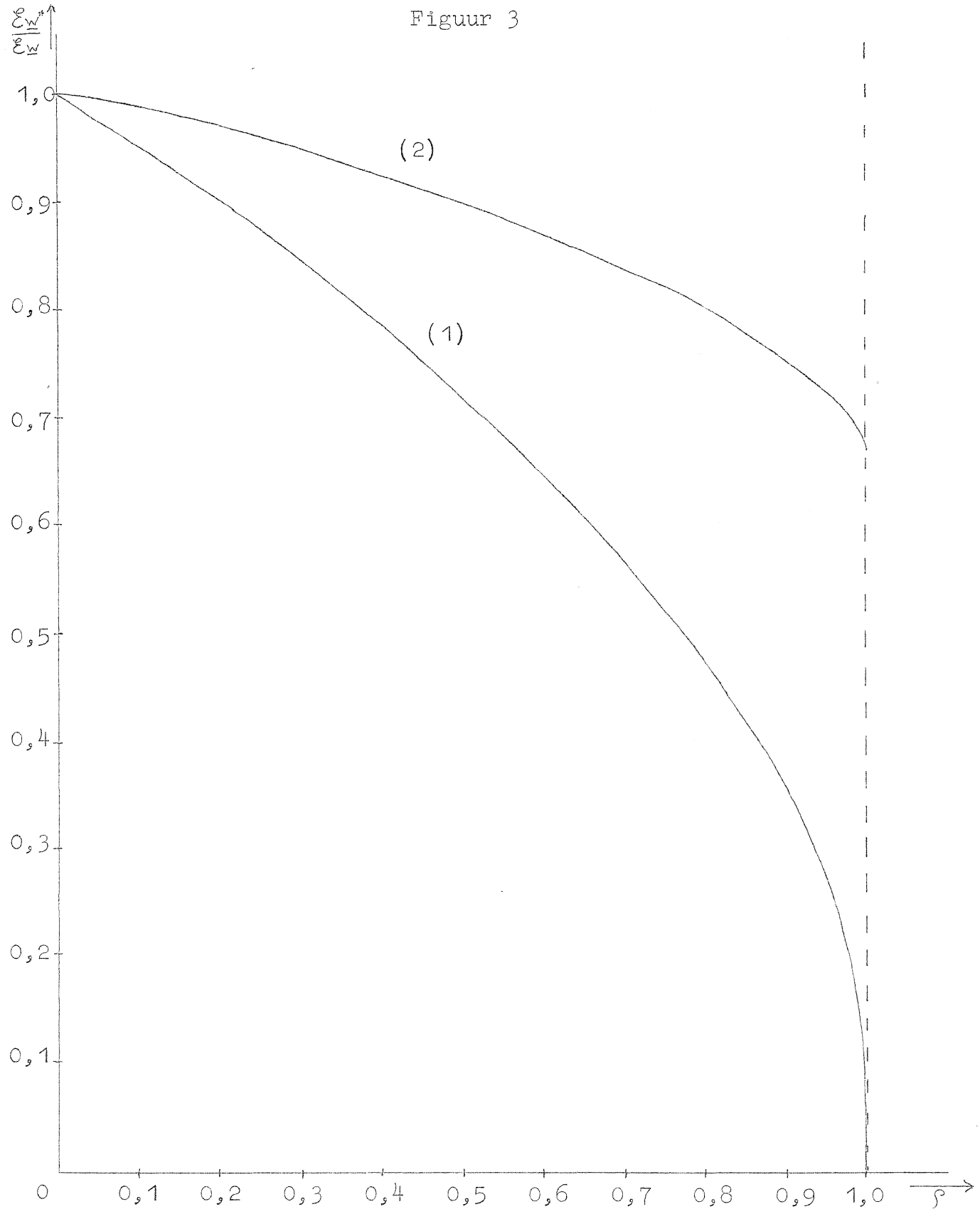
Vergelijking van figuur 2 en figuur 3 toont, dat in het

---

1) Nog steeds onder de voorwaarde, dat B(s) differentieerbaar is.

2) T.E. Phipps, Machine repair as a priority waiting-line problem, J.Op.Res. Soc. of America, 4, 76-86, (1956).

Figuur 3



$$\frac{1}{1-\rho}\int_0^{\infty}\frac{dB(s)}{\left\{1-\rho\int_0^{s}x\,dB(x)\right\}^2}$$

voor $B(s) = 1 - e^{-\frac{s}{\mu}}$ (1)

en $B(s) = \begin{cases} 0 & (s \leq \frac{1}{2}) \\ s - \frac{1}{2} & (\frac{1}{2} < s \leq \frac{3}{2}) \\ 1 & s > \frac{3}{2} \end{cases}$ (2)

laatste geval inderdaad een grotere vermindering van de gemiddelde wachttijd wordt bereikt, doch tevens, dat het grootste deel van de mogelijke vermindering van de gemiddelde wachttijd al optreedt bij een splitsing in twee groepen.

Daar bovendien de methode van Phipps tot vrij ingewikkelde berekeningen leidt en een nauwkeurige schatting van de bedieningstijden eist, zal het dikwijls aanbeveling verdienen, zich te beperken tot een splitsing in twee groepen. Kan men de bedieningstijden inderdaad nauwkeurig schatten, dan kan men de methode van Phipps toepassen.

Conclusies

De belangrijkste conclusies, die wij uit het voorgaande kunnen trekken, zijn

1) het verlenen van voorrang aan klanten met een kortere bedieningstijd voor klanten met een langere bedieningstijd geeft aanleiding tot een vermindering van de gemiddelde wachttijd.

2) bij een hogere bezettingsgraad wordt hierdoor een relatief grotere vermindering van de gemiddelde wachttijd bereikt.