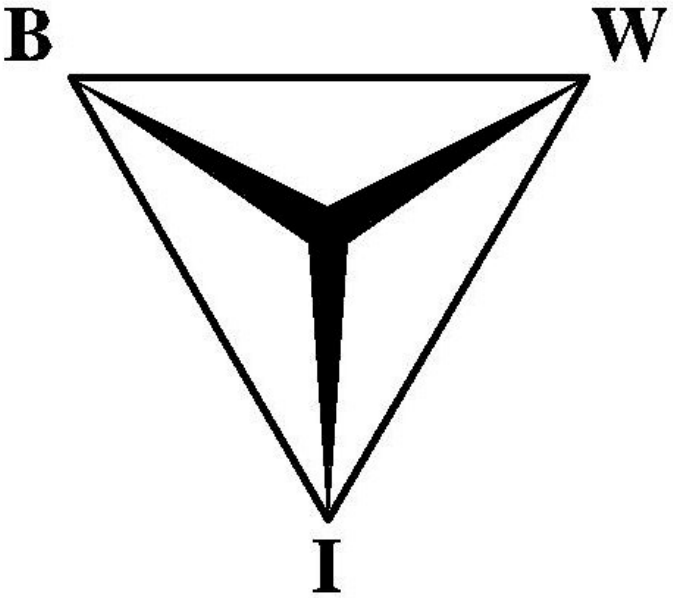
Risicobeheersing bij call center planning



BWI werkstuk - René van der Post

Vrije Universiteit
Faculteit der Exacte Wetenschappen
De Boelelaan 1081a
1081HV Amsterdam

Geschreven door:	René van der Post	-	rgvdpost@cs.vu.nl
Onder begeleiding van:	Ger Koole	-	koole@cs.vu.nl
	Auke Pot	-	sapot@cs.vu.nl

Inhoudsopgave

1 Inleiding

Voor u ligt het BWI werkstuk van René van der Post. Het BWI werkstuk is één van de laatste onderdelen van de de studie Bedrijfswiskunde en Informatica. Het doel is dat de student voor een deskundige manager een probleem beschrijft. Hierbij moet, naast de wiskunde en informatica aspecten, vooral nadruk worden gelegd op het bedrijfsgerichte aspect.

Ik heb voor mijn onderwerp gekozen omdat ik tijdens de colleges van *modeling of business processes* geïnteresseerd ben geraakt in call centers. In eerste instantie wilde ik me concentreren op de onzekerheid bij het voorspellen van het aantal benodigde agenten. In overleg met mijn begeleiders is later ook de onzekerheid wat betreft de aanwezigheid van personeel erbij betrokken.

Omdat dit werkstuk geschreven is voor verschillende doelgroepen, zijn er verschillende 'routes' om het te lezen. De call center manager zal waarschijnlijk geïnteresseerd zijn in de moeilijkheden en risico's die komen kijken bij het maken van de personeelsplanning. De wiskunde die hierachter schuilt zal hierbij minder van belang zijn aangezien dit veelal is voorgeprogrammeerd in de gebruikte beslissingsondersteunende systemen. Daarom zullen de hoofdstukken 4, 7, 8 en 9 voor de call center manager het interessantst zijn.

Voor de (bedrijfs)wiskundige die wil weten wat er zoal komt kijken bij het maken van de personeelsplanning in een call center, is ook hoofdstuk 3 opgenomen waarin kort wordt beschreven wat een call center eigenlijk is en hoe het doorgaans is opgebouwd. Deze lezer zal ook wat meer geïnteresseerd zijn in de gebruikte wiskundige methoden en wat minder in de implementatie hiervan in een beslissingsondersteunend systeem. De te volgen route wordt daarom de hoofdstukken 3, 4, 5, 6, 7, en 8.

Verder wordt er in hoofdstuk 2 een korte samenvatting van dit werkstuk gegeven en de belangrijkste conclusies worden op een rijtje gezet in hoofdstuk 10. Voor beide doelgroepen is de case studie (hoofdstuk 11) wellicht ook interessant omdat de in het werkstuk beschreven methoden hier aan de hand van voorbeelden duidelijk(er) gemaakt worden.
In de appendix worden beknopt alle gebruikte kansverdelingen en enkele van hun eigenschappen besproken.

Tot slot wil ik graag mijn begeleiders Ger Koole en Auke Pot bedanken voor hun hulp bij het tot stand komen van dit werkstuk.

2 Samenvatting

In dit werkstuk wordt besproken wat er gedaan kan worden om de heersende onzekerheid bij het maken van de planning te verkleinen. Hierbij wordt onderscheid gemaakt tussen twee situaties:

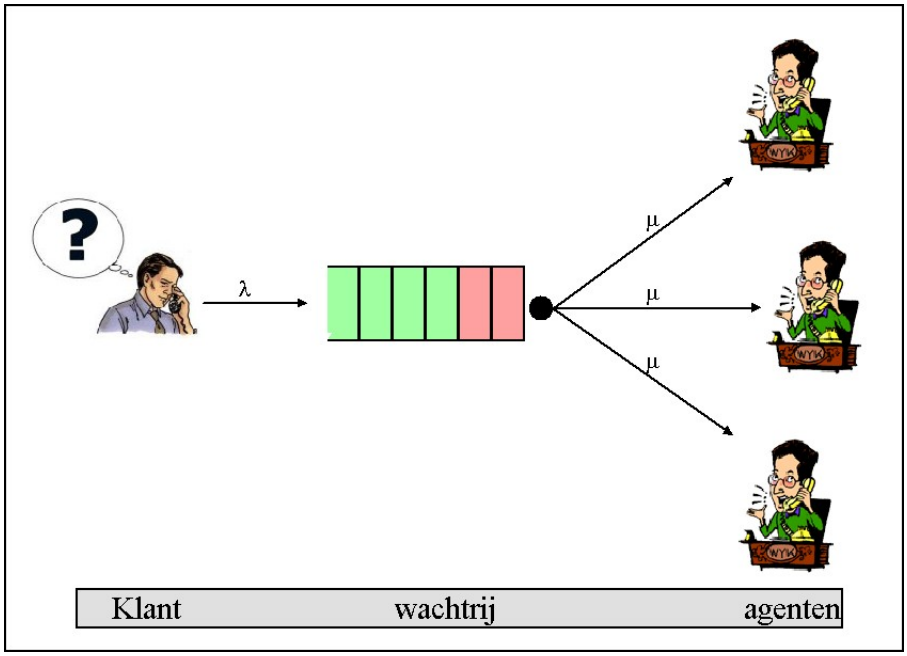
- ⇝ Ten eerste is er de onzekerheid bij het voorspellen van de hoeveelheid werk die verwacht kan worden. Hierbij is het vooral lastig om een betrouwbare schatting te maken van het aantal binnenkomende gesprekken. Met behulp van de in paragraaf 5.4 besproken Poisson mixtures methode kan hiervoor een betrouwbaarheidsinterval bepaald worden. De verwachte servicetijd is makkelijker te schatten. Daarom wordt hiervoor meestal het gemiddelde op basis van historische gegevens genomen. Aan de hand van de grenzen van het interval voor het aantal aankomsten en de verwachte servicetijd kan met de Erlang C formule (zie paragraaf 5.3) het aantal benodigde agenten bepaald worden.
- ⇝ Daarnaast is er ook geen zekerheid dat ingeroosterde agenten altijd de hele dag productief zijn; er zijn namelijk veel oorzaken waardoor werktijd verloren gaat. Dit verschijnsel heet shrinkage en wordt besproken in paragraaf 6.1. Om shrinkage tegen te gaan wordt er meestal een bepaald percentage agenten extra ingeroosterd. Hiervoor worden verschillende formules gebruikt, welke allemaal besproken worden in hoofdstuk 6.

Het gevolg van een foute voorspelling is dat er te veel of juist te weinig agenten aanwezig zijn in het call center. Dit geldt zowel bij een fout voorspelde aankomst- of servicerate als bij een fout geschat percentage shrinkage. Een tekort aan agenten leidt uiteraard tot slechte prestaties richting de klanten (lange wachttijden, slechte bereikbaarheid enz.). Een overschot zal over het algemeen leiden tot een te lage productiviteit en dus hoge kosten.

Er zijn verschillende manieren waarop onzekerheid bestreden kan worden. Zo kan er door zowel vaste als flexibele agenten in dienst te nemen vrij makkelijk geswitched worden naar een hoger aantal agenten als de situatie hierom vraagt (zie paragraaf 8.1). In paragraaf 8.2 wordt gekeken naar de mogelijkheid om agenten op verschillende afdelingen inzetbaar te maken; ook dit kan een positieve bijdrage leveren aan de prestaties van het call center. Door verschillende kanalen te bedienen, kan de productiviteit worden vergroot in geval van overbezetting. Een eventueel overschot aan agenten kan op rustige momenten bijvoorbeeld emails beantwoorden of uitgaande gesprekken voeren. Op drukke momenten kunnen alle agenten zich dan bezig houden met binnenkomende gesprekken (zie paragraaf 8.3 en 8.4).

3 Wat is een call center?

Laten we maar meteen beginnen met de definitie: een call center is een verzameling bronnen (voornamelijk agenten en ICT) die diensten levert per telefoon [9]. In zijn meest simpele vorm is een call center te zien als in het vogende plaatje:



Figuur 3.1

Er zijn s agenten (in dit geval geldt $s = 3$). Als een klant het call center belt, wordt er eerst gekeken of er een agent vrij is om het gesprek af te handelen. Als dit niet het geval is, dan neemt de klant plaats in de wachtrij totdat hij aan de beurt is. Bij het modelleren van een call center wordt er vaak vanuit gegaan dat er oneindig veel plaats is in de wachtrij en dat klanten altijd zullen blijven wachten totdat ze aan de beurt zijn. In werkelijkheid is dit natuurlijk niet het geval; de aanname van een oneindige wachtrij is al niet erg realistisch omdat er maar een beperkt aantal telefoonlijnen beschikbaar is. Als een klant belt terwijl alle lijnen bezet zijn, zal de klant geblokkeerd worden en niet in de wachtrij plaats kunnen nemen.
Ook zal een klant in werkelijkheid niet oneindig veel geduld hebben, na een bepaalde periode is het geduld op en zal de klant de verbinding verbreken (om het wellicht op een later tijdstip nog eens proberen).

Ook de opbouw van een call center is over het algemeen iets ingewikkelder dan in figuur 3.1, zo kan er bijvoorbeeld nog onderscheid gemaakt worden tussen:

- ⇝ *binnenkomende of uitgaande gesprekken*
  Er zijn twee mogelijkheden waarop een gesprek tot stand kan komen: de klant heeft een probleem en belt naar het call center (binnenkomend gesprek) of de agent belt naar de (potentiële) klant (uitgaand gesprek). De meeste call centers hebben zowel binnenkomende als uitgaande gesprekken, hierover meer in paragraaf 8.3.
- ⇝ *verschillende vaardigheden*
  Zeker naarmate een call center groter wordt, zullen er groepen agenten zijn met elk verschillende vaardigheden. Een voorbeeld hiervan is een call center van een mutinational waar klanten in verschillende talen te woord gestaan moeten worden. Voor iedere taal is er dan een groep agenten, of er zijn agenten die meerdere talen spreken (multiple skills). Als alle agenten slechts één taal spreken, is er eigenlijk sprake van een aantal kleine call centers binnen het call center. Bij multiple skills wordt de analyse wat lastiger.

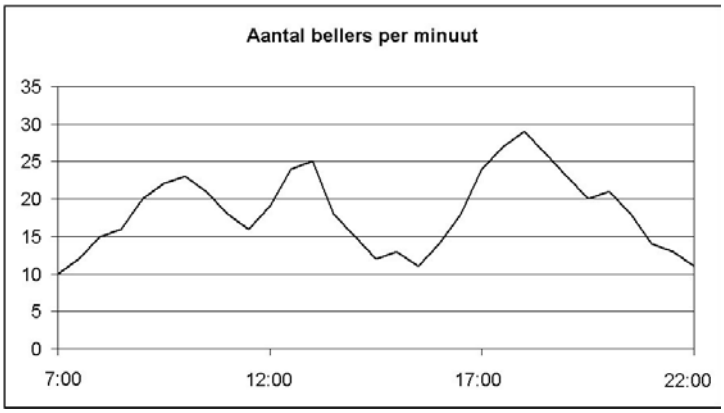
Bij het modelleren van een call center zal ik in eerste instantie, vanwege de eenvoud, uitgaan van de meest simpele vorm. Later zal ik dit uitbreiden naar een wat realistischer model.

Om de prestaties van een call center te meten wordt er meestal gekeken naar het zogenaamde servicelevel. Een veel gebuikte definitie van het servicelevel is dat $\alpha\%$ van de klanten binnen β seconden geholpen moet worden. Vaak wordt er gekozen voor $\alpha = 80\%$ en $\beta = 20$ seconden (een 80/20 servicelevel).
Andere prestatiemaatstaven voor het call center zijn bijvoorbeeld het percentage klanten dat voortijdig afhaakt, de gemiddelde wachttijd, de productiviteit van de agenten en natuurlijk de (loon)kosten.

4 Voorspellen, dat is lastig!

Om een goede planning te kunnen maken is het noodzakelijk om te weten hoeveel werk er (ongeveer) verwacht kan worden. Aan de hand van historische gegevens kan hiervan een schatting gemaakt worden. Het is echter erg lastig om een goede schatting te maken. Oorzaken hiervan zijn onder andere:

- Het is niet erg realistisch om het aantal bellers per minuut als constante te zien gedurende de dag (zie figuur 4.1).



Figuur 4.1

  We zien hier dat er op een dag een paar pieken (bijvoorbeeld rond 18:00 uur) en dalen ('s ochtends vroeg) zijn. Daarom wordt de dag meestal opgedeeld in kleine intervallen van bijvoorbeeld een half uur. Voor ieder interval wordt vervolgens de aankomstrate geschat.
  Voor de servicetijden (de servicerate) is een dergelijke opdeling minder noodzakelijk omdat deze minder afhankelijk is van het moment van de dag. Hierdoor is er meestal ook meer relevante data over beschikbaar waardoor er nauwkeuriger geschat kan worden.

- In theorie blijkt het servicelevel erg gevoelig te zijn voor kleine schommelingen in de voorspelling. Daarom zouden de voorspellingen erg nauwkeurig moeten zijn om een slecht servicelevel (of een te lage productiviteit) zo veel mogelijk te voorkomen. In de praktijk zijn er wel wat manieren waarop de prestaties van het call center minder gevoelig gemaakt kunnen worden voor foute voorspellingen. Deze zal ik bespreken in hoofdstuk 8.

- De werkelijkheid is van erg veel factoren afhankelijk, bijvoorbeeld het uur van de dag, de dag van de week, de maand van het jaar enz. Ook het weer kan een rol spelen. Met veel van deze factoren (bijvoorbeeld het weer) kan geen rekening gehouden worden bij het doen van een voorspelling omdat informatie hierover niet of veel te laat bekend is.

- ⇝ Er zijn veel afhankelijkheden. Zo blijkt het aantal binnenkomende gesprekken in verschillende perioden sterk van elkaar afhankelijk te zijn. Dit houdt in dat als het 's ochtends vroeg al drukker is dan verwacht, het zeer waarschijnlijk is dat het de hele dag drukker zal worden dan vooraf werd verwacht [7].
- ⇝ Bepaalde maatregelen binnen het bedrijf kunnen invloed hebben op het aantal binnenkomende gesprekken. Zo zal een verhoogd marketingbudget ongetwijfeld leiden tot een toename van het aantal gesprekken, er kan dan immers meer reclame gemaakt worden voor een bepaald product of dienst waardoor er als het goed is meer vraag naar komt en er dus meer mensen op vragen of problemen stuiten. Daarom moet er goed overlegd worden tussen de verschillende afdelingen en moet er voor gezorgd worden dat de juiste informatie op de juiste plek terecht komt.
- ⇝ Data is vaak niet meer (goed) bruikbaar omdat er bepaalde veranderingen binnen het call center kunnen hebben plaatsgevonden, bijvoorbeeld een andere (efficiëntere) manier van toekennen van telefoongesprekken aan agenten. Historische data is dan mogelijk niet meer relevant genoeg om nog goede schattigen voor de nieuwe situatie te kunnen maken.

Ondanks al deze moeilijkheden moet de planner toch een redelijk goede voorspelling zien te maken...

Figuur 4.2

5 Hoeveel agenten roosteren?

Bij het bepalen van de hoeveelheid benodigde agenten spelen twee factoren een belangrijke rol, namelijk de aankomstrate en de servicerate.

5.1 Voorspellen van de aankomstrate

De aankomstrate (meestal aangeduid met λ) geeft de snelheid waarmee telefoongesprekken binnenkomen bij het call center aan en is gedefinieerd als $\lambda = 1/\textit{aantal aankomsten per tijdseenheid}$. Zoals gezegd wordt de dag meestal opgedeeld in kleine intervallen en wordt voor ieder interval de aankomstrate bepaald. Het bepalen van deze aankomstrate is echter moeilijk omdat er veel externe factoren een rol bij spelen. Zo kan bijvoorbeeld het weer of de introductie van een nieuw product de aankomstrate beïnvloeden. Omdat het zo lastig is om de aankomstrate te schatten, kan er ook gewerkt worden met een $100(1-2\alpha)\%$ betrouwbaarheidsinterval [11]

$$\lambda \in \left[\hat{\lambda} - \xi_{1-\alpha}\sqrt{\hat{\lambda}/K}, \hat{\lambda} + \xi_{1-\alpha}\sqrt{\hat{\lambda}/K}\right]$$

Hierin is $\hat{\lambda}$ de gemiddelde aankomstrate, K is het aantal waarnemingen en ξ_β is het β-kwantiel van de standaard normale verdeling (dus $\Phi(\xi_\beta) = \beta$).
Meestal wordt er gewerkt met het 95% betrouwbaarheidsinterval, de waarde van $\xi_{0.975}$ is dan ongeveer 1.96

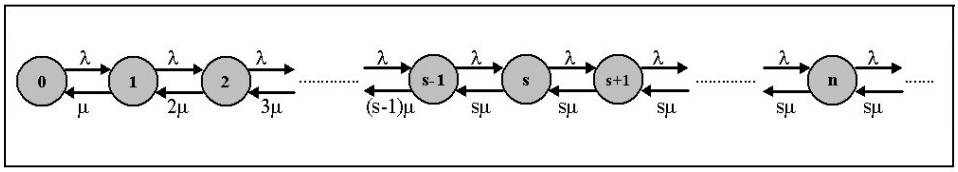
5.2 Voorspellen van de servicerate

De servicerate (μ) representeert de snelheid waarmee gesprekken afgehandeld worden. Omdat de servicerate minder afhankelijk is van externe factoren dan de aankomstrate en omdat er vaak meer relevante data over beschikbaar is, wordt er meestal met een constante servicerate gewerkt. Hiervoor wordt dan de puntschatter op basis van historische gegevens gebruikt.
Overigens gaat het hier niet alleen om de tijd dat de agent daadwerkelijk in gesprek is met de klant. Ook het raadplegen of bijwerken van dossiers voor of na het gesprek moet bij de servicetijd gerekend worden.

5.3 De Erlang formule

Nu de aankomst- en servicerate bepaald zijn, kunnen we met behulp van de zogenaamde Erlang formule uitrekenen hoeveel agenten (s) we nodig denken te hebben om een bepaald servicelevel te bereiken.
Als we veronderstellen dat de aankomst- en servicetijden exponentieel verdeeld zijn en dat het systeem in een stationaire toestand verkeert, kunen we deze Erlang formule afleiden. Hiervoor moet dan wel gelden dat $\lambda/s\mu < 1$ omdat het systeem anders instabiel is waardoor het aantal klanten in het systeem naar oneindig zal gaan.

Zoals gezegd veronderstellen we ook dat er oneindig veel plaats is in de wachtrij en dat klanten oneindig veel geduld hebben. We hebben het call center nu gemodelleerd als een $M/M/s/\infty$ rij en kunnen het dus analyseren met behulp van onderstaande Markov-keten:



Figuur 5.1

Hiermee bepalen we nu de evenwichtsvergelijkingen voor de stationaire toestanden (π_n):

$$\begin{array}{lll}
\lambda\pi_0 = \mu\pi_1 & \Rightarrow \pi_1 & = \frac{\lambda}{\mu}\pi_0 \\
\lambda\pi_1 = 2\mu\pi_2 & \Rightarrow \pi_2 & = \frac{\lambda}{2\mu}\pi_1 \\
 & & = \frac{1}{2}\left(\frac{\lambda}{\mu}\right)^2\pi_0 \\
\lambda\pi_2 = 3\mu\pi_3 & \Rightarrow \pi_3 & = \frac{\lambda}{3\mu}\pi_2 \\
 & & = \frac{1}{6}\left(\frac{\lambda}{\mu}\right)^3\pi_0 \\
\ldots & & \ldots \\
\lambda\pi_{n-1} = n\mu\pi_n & \Rightarrow \pi_n & = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n\pi_0
\end{array}$$

Als we nu $a = \lambda/\mu$ substitueren, vinden we dus dat voor een willekeurige π_n met $n < s$ geldt dat $\pi_n = \frac{a^n}{n!}\pi_0$. Op dezelfde manier kunnen we voor $n \geq s$ vinden dat $\pi_n = \frac{a^n}{s!s^{n-s}}\pi_0$.
Verder weten we dat $\sum_{n=0}^{\infty}\pi_n = 1$ en dus kunnen we nu π_0 bepalen:

$$\begin{array}{rl}
\pi_0 & = \left[\sum_{n=0}^{s-1}\frac{a^n}{n!} + \sum_{n=s}^{\infty}\frac{a^n}{s!s^{n-s}}\right]^{-1} \\
 & = \left[\sum_{n=0}^{s-1}\frac{a^n}{n!} + \frac{s^s}{s!}\sum_{n=s}^{\infty}\left(\frac{a}{s}\right)^n\right]^{-1} \\
 & = \left[\sum_{n=0}^{s-1}\frac{a^n}{n!} + \frac{s^s}{s!}\frac{(a/s)^s}{1-(a/s)}\right]^{-1} \\
 & = \left[\sum_{n=0}^{s-1}\frac{a^n}{n!} + \frac{sa^s}{s!(s-a)}\right]^{-1} \\
 & = \left[\sum_{n=0}^{s-1}\frac{a^n}{n!} + \frac{a^s}{(s-1)!(s-a)}\right]^{-1}
\end{array}$$

Bij de derde stap maken we gebruik van het feit dat we te maken hebben met een meetkundige reeks. Omdat we nu de uitdrukking voor π_0 weten, kunnen we vervolgens alle andere π_n's uitrekenen.

Nu zijn we geïnteresseerd in de kansverdeling van de wachttijd (W). Hiervoor hebben we eerst de kans nodig dat er überhaubt gewacht moet worden, ofwel $\mathbb{P}(W > 0)$. Deze kans wordt ook wel aangeduid met $C(s, a)$. Door de PASTA-eigenschap (Poisson Arrivals See Time Averages, zie [15], paragraaf 9.3) is dit gelijk aan de fractie tijd dat er s of meer klanten in het systeem zijn, we vinden hiervoor:

$$\begin{aligned} C(s,a) &= \sum_{n=s}^{\infty} \pi_n \\ &= \sum_{n=s}^{\infty} \frac{a^n}{s!s^{n-s}} \left[\sum_{n=0}^{s-1} \frac{a^n}{n!} + \frac{a^s}{(s-1)!(s-a)} \right]^{-1} \\ &= \frac{a^s}{(s-1)!(s-a)} \left[\sum_{n=0}^{s-1} \frac{a^n}{n!} + \frac{a^s}{(s-1)!(s-a)} \right]^{-1} \end{aligned}$$

Merk op dat we hier gebruik maken van dezelfde meetkundige reeks als bij het afleiden van π_0.

Nu bepalen we wat de verdeling is van de wachttijd, als er gegeven is dat de wachttijd groter is dan nul. We definiëren hiervoor eerst ρ als $\rho = a/s$.

$$\begin{aligned} \mathbb{P}(W > t | W > 0) &= \sum_{n=0}^{\infty} \mathbb{P}(W > t | L = n + s, W > 0) \mathbb{P}(L = n + s, W > 0) \\ &= \sum_{n=0}^{\infty} \mathbb{P}(W > t | L = n + s) \mathbb{P}(L = n + s, W > 0) \\ &= (1-\rho) e^{-s\mu t} \sum_{j=0}^{\infty} \sum_{k=0}^{j} \frac{(s\mu t)^k}{k!} \rho^j \\ &= (1-\rho) e^{-s\mu t} \sum_{k=0}^{\infty} \frac{(s\mu t)^k}{k!} \sum_{j=k}^{\infty} \rho^j \\ &= e^{-s\mu t} \sum_{k=0}^{\infty} \frac{(\rho s\mu t)^k}{k!} \\ &= e^{-(1-\rho)s\mu t} \\ &= e^{-(s\mu - \lambda)t} \end{aligned}$$

Bij de derde stap maken we gebruik van het feit dat $\mathbb{P}(W > t | L = n + s)$ gelijk is aan de kans dat er minder dan $(n + 1)$ vertrekken zijn in een periode met lengte t. Het aantal vertrekken heeft een Poisson verdeling met parameter $(s\mu t)$ [14] dus als V het aantal vertrekken is, dan is
$\mathbb{P}(V = k) = e^{-s\mu t} \cdot (s\mu t)^k / k!$. Verder heeft de rijlengte, gegeven het feit dat er een rij is, een geometrische verdeling met parameter ρ [8] en dus als L de rijlengte is, dan geldt dat $\mathbb{P}(L = s + j | W_Q > 0) = \mathbb{P}(L = n | L > s) = (1 - \rho)\rho^j$.

Nu we de uitdrukkingen voor $\mathbb{P}(W > 0)$ en $\mathbb{P}(W > t | W > 0)$ weten, kunnen we eindelijk bepalen wat $\mathbb{P}(W > t)$ is, we vinden hiervoor:

$$\begin{aligned} \mathbb{P}(W > t) &= \mathbb{P}(W > 0) \mathbb{P}(W > t | W > 0) \\ &= C(s,a) e^{-(s\mu - \lambda)t} \end{aligned}$$

We kunnen nu het servicelevel definiëren als $SL = 1 - \mathbb{P}(W > t)$. Met deze formule kunnen we berekenen hoeveel agenten er nodig zijn om bijvoorbeeld een 80/20 servicelevel te halen. De formule staat bekend als de Erlang C formule en is veelal geïmplementeerd in beslissingsondersteunende systemen (zie bijvoorbeeld www.math.vu.nl/∼koole/ccmath/ErlangC.php).

Stel dat we nu met het $100(1-\alpha)\%$ betrouwbaarheidsinterval voor λ hadden gewerkt, dan kunnen we hiermee ook het $100(1-\alpha)\%$ betrouwbaarheidsinterval voor het aantal benodigde agenten bepalen, door simpelweg de onder- en bovengrens van het interval in te vullen.

In de Erlang C formule worden twee belangrijke aspecten buiten beschouwing gelaten. Allereerst wordt er geen rekening gehouden met het blokkeren van binnenkomende gesprekken. In werkelijkheid is er namelijk maar een beperkt aantal lijnen beschikbaar en als deze allemaal bezet zijn, worden binnenkomende gesprekken geblokkeerd. Ook gaan we er vanuit dat bellers een oneindige hoeveelheid geduld hebben. In werkelijkheid is dit natuurlijk niet zo, veel mensen zullen ophangen als ze na een bepaalde periode (gemiddeld ongeveer één minuut) nog niet geholpen zijn.
Als we deze twee factoren wel meenemen in het model, krijgen we de zogenaamde Erlang X formule, welke ik hier niet zal afleiden. Ook deze formule kan geïmplementeerd worden in een beslissingsondersteunend systeem zoals bijvoorbeeld op www.math.vu.nl/∼koole/ccmath/ErlangX.php.

5.4 Een betere manier?

In de praktijk blijkt de aanname van exponentieel verdeelde aankomsttijden niet erg realistisch te zijn: als dit wel het geval zou zijn, zou het aantal aankomsten in een bepaalde periode Δt namelijk een Poisson$(\lambda\Delta t)$ verdeling hebben [14]. De verwachting en de variantie van een Poisson(λ) verdeelde variabele X zouden hetzelfde moeten zijn, namelijk $\mathbb{E}X = \mathrm{Var}X = \lambda$. In de praktijk blijkt de variantie echter veel groter te zijn dan het gemiddelde [8].
Daarom kan het beter zijn om te werken met een zogenaamd *gemixed Poisson model*. Het genereren van de aankomsten gebeurt hierbij in twee stappen:

- ⇝ We trekken de aankomstrate λ voor een bepaalde periode uit een kansverdeling met verdelingsfunctie H,
- ⇝ Met de getrokken λ trekken we een Poisson verdeelde variabele X als aantal aankomsten in die periode.

De verdelingsfunctie van X is dan $\mathbb{P}_H(X = x) = \int_0^\infty \frac{\lambda^x}{x!} e^{-\lambda} dH(\lambda)$.
Als we nu een reeks realisaties $x_1, \ldots x_n$ hebben, willen we aan de hand daarvan de verdelingsfunctie H bepalen. Aan de hand van deze H zouden we dan een onder- en bovengrens voor λ kunnen bepalen, zo dat
$\mathbb{P}_H(\lambda \in [\lambda^-, \lambda^+]) \geq 1 - 2\alpha$. Het mooiste zou zijn als we precies de verdelingsfunctie H wisten, we zouden dan immers de kwantielen $H^{-1}(\alpha)$ en $H^{-1}(1-\alpha)$ kunnen nemen als onder- respectievelijk bovengrens voor λ. Helaas weten we H in de praktijk niet en zullen we deze dus moeten schatten. Dit kan op verschillende manieren, waarvan ik er twee zal bespreken.

Parametrische schatting .
Bij deze methode wordt er een kansverdeling gekozen en worden vervolgens aan de hand van de realisaties de parameters behorende bij deze verdeling geschat. Het voordeel van deze methode is dat de parameters makkelijk te schatten zijn. Als we bijvoorbeeld een Gamma(r,s) verdeling kiezen, dan zien we dat $h_{r,s}(\lambda) = \frac{s^r}{\Gamma(r)}\lambda^{r-1}e^{-\lambda s}1_{[0,\infty]}$ en kunnen we voor $\mathbb{P}_{r,s}(X = k)$ vinden dat:

$$\mathbb{P}_{r,s}(X = k) = \binom{r+k-1}{k}\left(\frac{s}{1+s}\right)^r\left(1 - \frac{s}{1+s}\right)^k$$

Dit is een negatief binomiale verdeling met succeskans $s/(1+s)$ en r successen [8]. We kunnen nu met behulp van de verwachting en variantie $\hat{r}$ en $\hat{s}$ te schatten. Hiervoor vinden we:

$$\hat{r} = \frac{(\mathbb{E}X)^2}{\mathbb{E}X + \mathrm{Var}X} \text{ en } \hat{s} = \frac{\mathbb{E}X}{\mathrm{Var}X}$$

Met behulp van deze schattingen voor $\hat{r}$ en $\hat{\lambda}$ kunnen we de kwantielen $H^{-1}_{\hat{r},\hat{s}}(\alpha)$ en $H^{-1}_{\hat{r},\hat{s}}(1-\alpha)$ bepalen. Deze kwantielen kunnen we vervolgens gebruiken als onder- respectievelijk bovengrens voor λ.

Non-parametrische schatting .
Een andere manier om H te bepalen is een zogenaamde non-parametrische methode, bijvoorbeeld met behulp van de meest aannemelijke schatter. Stel dat we een verzameling realisaties $x_1, \ldots, x_n$ hebben, we zijn dan op zoek naar de dichtheid van $X = (X_1, \ldots, X_n)$ ofwel $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$. Hiertoe introduceren we als aannemlijkheidsfunctie $\theta \to L(\theta, x) := p_\theta(x)$, waarin $p_\theta(x)$ de kansdichtheid van X_i is. De meest aannemelijke schatter voor θ is dan de waarde van θ die de aannemelijkheidsfunctie maximaliseert.
Omdat de realisaties gelijk verdeeld en onderling onafhankelijk zijn, is de dichtheid van X gelijk aan het product $\prod_{i=1}^n p_\theta(x_i)$ en omdat we alleen geïnteresseerd zijn in de plaats van het maximum van de aannemelijkheidsfunctie $L(\theta, x_1, \ldots, x_n)$ en niet in de grootte ervan, kunnen we ook het logaritme hiervan nemen, we vinden dan:

$$\log L(\theta, x_1, \ldots, x_n) = \textstyle\sum_{i=1}^n \log p_\theta(x_i)$$

Nu vullen we voor $p_\theta(x)$ de formule voor $\mathbb{P}_H(X = x)$ in:

$$\log L(\theta, x_1, \ldots, x_n) = \sum_{j=0}^{k} m_j \log \int_0^\infty \lambda^j e^{-\lambda} dH(\lambda) + C$$

Hierin is $k = \max_i x_i$ en m_j is het aantal keren dat j voorkomt in de realisatie $x_1, \ldots, x_k$. [11]

Nu moeten we de volgende vergelijking oplossen om de meest aannemelijke schatter te vinden:

$$\frac{\delta}{\delta\theta_j} \log L(\theta, x) = 0$$

Deze vergelijking is vaak niet expliciet op te lossen. Wel zijn er verschillende algoritmes waarmee een oplossing gevonden kan worden, bijvoorbeeld met het programma C.A.MAN (Computer Assisted Mixture ANalysis: zie http://www.medizin.fu-berlin.de/sozmed/caman.html).

5.5 Vergelijking

Omdat er in paragraaf 5.3 kennelijk verkeerde aannames zijn gemaakt, namelijk dat de aankomsttijden exponentieel verdeeld zijn, zegt mijn gevoel me dat de tweede manier om de aankomstrate te schatten (Poisson mixtures) wat betrouwbaarder is.

In de volgende tabel zijn de intervallen voor het aantal aankomsten per minuut volgens de twee methoden te zien: in de tweede kolom is het interval geschat met behulp van de in paragraaf 5.1 gebruikte methode, in de derde kolom is de in paragraaf 5.4 besproken methode gebruikt. De waarden zijn gevonden met behulp van de data van Avi Mandelbaum (zie http://iew3.technion.ac.il/serveng/).

Interval	Paragraaf 5.1	Poisson mixtures
7:00- 7:30	[0.343, 0.444]	[0.068, 0.622]
7:30- 8:00	[0.536, 0.660]	[0.132, 0.974]
8:00- 8:30	[0.727, 0.870]	[0.207, 1,281]
8:30- 9:00	[0.898, 1.056]	[0.309, 1,191]
9:00- 9:30	[0.936, 1.097]	[0,328, 1.221]
9:30-10:00	[1.132, 1.308]	[0.396, 1,846]
10:00-10:30	[1.198, 1.379]	[0.460, 1,579]
10:30-11:00	[1.198, 1.379]	[0.429, 1.935]
11:00-11:30	[1.127, 1.303]	[0.418, 1.605]

Tabel 5.1

We zien dat de schatting van paragraaf 5.1 en de Poisson mixtures methode ongeveer dezelfde verwachte aankomstrate geven, alleen is de spreiding bij Poisson mixtures veel groter. Blijkbaar is het ook nodig om rekening te houden met deze grotere spreiding van de aankomsten, uit onderzoek blijkt immers dat het aantal aankomsten niet Poisson verdeeld is zoals meestal aangenomen wordt. Daarom denk ik dat de Poisson mixtures methode een wat realistischer beeld geeft van de werkelijkheid.

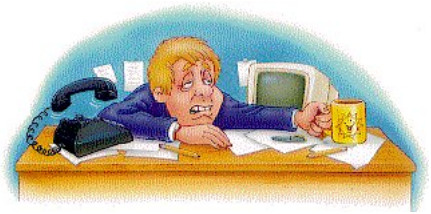
Om echter een goed gefundeerd oordeel te geven over welke methode beter is, zullen er volgens mij meer praktijksituaties onderzocht moeten worden: er is nu één dataset van één call center onderzocht en op basis daarvan is mijn keuze waarschijnlijk niet gerechtvaardigd...

6 Een zieke agent...

6.1 Shrinkage

Als het aantal benodigde agenten eenmaal bepaald is, zijn we er nog niet: we hebben immers geen zekerheid dat iedere agent die is ingeroosterd ook daadwerkelijk (op tijd) op komt dagen. Hij kan namelijk ziek zijn, in de file staan enz. Ook als de agenten eenmaal aanwezig zijn, zijn ze niet iedere minuut van de dag in staat om klanten te helpen. Zo kan een agent bijvoorbeeld naar het toilet zijn, met collega's overleggen enz. Dit soort 'verloren tijd' wordt ook wel *shrinkage* genoemd. Andere vormen van shrinkage zijn bijvoorbeeld [12]:

- ⇝ (lunch)pauzes
- ⇝ vergaderingen
- ⇝ afwezigheid (o.a. ziekte)
- ⇝ onderzoek



Figuur 6.1

Er wordt geen rekening gehouden met het optreden van shrinkage bij het bepalen van het aantal agenten dat nodig is om het gewenste servicelevel te bereiken (de basis staf). Om er toch voor te zorgen dat het servicelevel niet te laag wordt, wordt deze basis staf over het algemeen verhoogd met een vooraf bepaald percentage. Meestal wordt hiervoor een percentage gebruikt van rond de 30% maar uitschieters naar 15% of 50% zijn geen uitzondering. Het verhogen van de basis staf kan gebeuren op twee manieren: de lineaire en de inverse methode. Beide methoden zullen zo dadelijk worden besproken.

Het schatten van het percentage shrinkage gebeurt aan de hand van historische data. Hierbij moet wel goed opgelet worden wat er als shrinkage gezien wordt en wat niet. In sommige call centers wordt bijvoorbeeld het bijwerken van dossiers na een gesprek gezien als shrinkage en bij andere call centers wordt deze tijd weer bij de servicetijd gerekend. Er moet dus wel worden voorkomen dat deze 'dossiertijd' dubbel (of juist helemaal niet) meegeteld wordt. Verder is het misschien beter om bij het schatten van het percentage shrinkage onderscheid te maken tussen shrinkage waarvan je zeker kunt weten dat het zal optreden (zoals bijvoorbeeld vergaderingen) en shrinkage waarvan je die zekerheid niet hebt (bijvoorbeeld ziekte). Het ziekteverzuim ligt gemiddeld rond de 10%, maar ik vermoed dat de variantie vrij groot zal zijn.

Ook zijn er ongetwijfeld bepaalde dagen dat het ziekteverzuim met zekerheid hoger zal zijn dan gemiddeld (de vrijdag voor Pinksterweekend...). Het heeft daarom volgens mij niet zoveel zin om altijd uit te gaan van het gemiddelde: stel dat er in een call center waar twintig agenten nodig zijn, er gemiddeld twee ziek zijn. Als je dan iedere dag twee agenten extra inroostert maar van maandag tot en met donderdag zijn er geen zieken en op vrijdag zijn er tien, dan kom je weliswaar gemiddeld op twee zieken per dag uit, maar van maandag tot en met donderdag heb je twee agenten teveel en op vrijdag heb je er alsnog acht te weinig. Daarom denk ik dat het beter is om ziekteverzuim niet mee te nemen in het percentage shrinkage. Eventuele ziektegevallen kunnen opgevangen worden met behulp van flexibele agenten die gebeld worden zodra dat nodig is.

Een andere optie is om wel een bepaald (laag) percentage agenten extra in te roosteren in verband met mogelijk ziekteverzuim en een eventueel hoger percentage op te vangen met flexibele agenten. In een call center met honderd agenten zal een dag zonder zieken namelijk vrij uitzonderlijk zijn en zullen er altijd wel een paar agenten ziek zijn. Dit houdt dus in dat als ziektepercentage niet het gemiddelde maar bijvoorbeeld het minimum of het gemiddelde min de standaarddeviatie gebruikt moet worden.

Als flexibele agenten kunnen bijvoorbeeld studenten gebruikt worden. Kijkend naar mijn eigen situatie zou ik, als ik op dit moment gebeld zou worden met de vraag of ik zou kunnen komen werken omdat er een zieke is, makkelijk kunnen komen (als ik ten minste in een call center zou werken). Het schrijven van dit werkstuk zou ik dan wel kunnen uitstellen tot vanavond of morgen...

Op zich is het ook vreemd om bijvoorbeeld vergaderingen mee te nemen bij het berekenen van het percentage shrinkage. Deze worden namelijk ruim van tevoren gepland en er is ongetwijfeld bekend welke agenten hierbij aanwezig zullen moet zijn. Deze agenten moeten voor de betreffende tijdstippen dan gewoon niet ingeroosterd worden voor hun normale werkzaamheden. Hetzelfde geldt voor (lunch)pauzes: er is van tevoren bekend dat en wanneer elke agent pauze moet houden. Als de werk- en pauzetijden duidelijk gedefinieerd worden, zou ook dit niet meegenomen hoeven worden in het percentage shrinkage. Door het lijstje verder uit te kleden kunnen we denk ik een percentage shrinkage bepalen wat enkel bestaat uit het ziekteverzuim en eventueel een paar procent non-productieve tijd zoals bijvoorbeeld toiletbezoek.

Een dergelijke aanpak vergt misschien wat meer administratief werk; de werktijden moeten immers nauwkeurig gedefinieerd worden en afwezigheid van agenten door vergaderingen, vakanties enz. moeten goed worden bijgehouden. Ook zal het maken van het rooster waarschijnlijk wat gecompliceerder worden.
Als dit alles echter goed geregeld is, dan zal dit volgens mij leiden tot betere resultaten dan de manier waarop het nu meestal gebeurt.

Overigens kan ik me haast niet voorstellen dat dit niet de normale gang van zaken is in call centers, in alle literatuur die ik heb bestudeerd staat echter dat vergaderingen, pauzes enz. ook in de berekening van het percentage shrinkage meegenomen moeten worden.

Merk op dat het ook waarschijnlijk niet altijd mogelijk is om bij het plannen rekening te houden met vergaderingen: als planningen lang van tevoren gemaakt worden en vergaderingen op korte termijn gepland worden, kan hier uiteraard geen rekening mee gehouden worden.

6.2 De lineaire methode

Bij deze methode wordt de basis staf verhoogd met behulp van de formule:

$$\textit{benodigde staf} = \textit{basis staf} \cdot (1 + \%\ \textit{shrinkage})$$

Stel dat we een percentage shrinkage hanteren van 30% en we met behulp van een Erlang C calculator een basis staf van dertig agenten hebben gevonden, dan besluiten we dus om $30 \cdot (1 + 0.3) = 39$ agenten in te roosteren.
Bij deze methode gaat men er van uit dat alleen de verloren tijd van de basis staf opgevuld hoeft te worden door extra agenten. Met andere woorden: als er op een dag acht uur shrinkage is, dan wordt dit gat opgevuld met één extra agent.

6.3 De inverse methode

Ook bij deze methode wordt er gewerkt met een vast percentage shrinkage, maar in tegenstelling tot bij de lineaire methode wordt hier ook rekening mee gehouden met het feit dat extra agenten ook weer zullen leiden tot extra shrinkage. Als we de benodigde staf definiëren als s_n, de basis staf als s_b en het percentage shrinkage als σ dan wordt de formule om de benodigde staf te berekenen:

$$\begin{aligned} s_n &= s_b + \sigma s_b + \sigma^2 s_b + \ldots + \sigma^n s_b + \ldots \\ &= s_b \textstyle\sum_{i=0}^{\infty} \sigma^i \end{aligned}$$

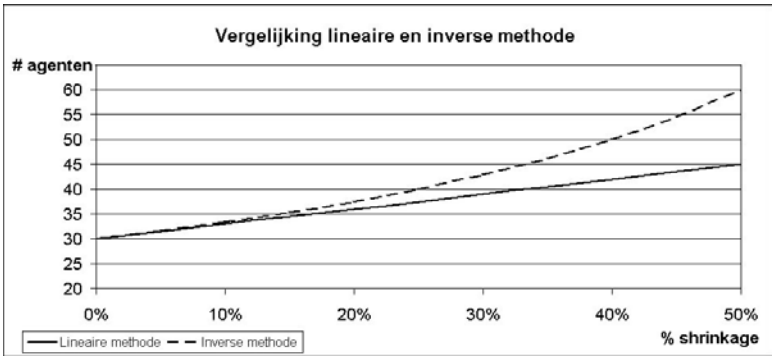
Dit is een meetkundige reeks en omdat $\sigma < 1$, convergeert deze naar $s_b/(1-\sigma)$. Uiteindelijk vinden we dus:

$$\textit{benodigde staf} = \frac{\textit{basis staf}}{1 - \%\ \textit{shrinkage}}$$

In het voorbeeld uit de vorige paragraaf zullen we nu dus besluiten om $30/(1-0.3) \approx 43$ agenten in te roosteren.
Merk op dat er, omdat er hier ook altijd rekening gehouden wordt met het optreden van shrinkage bij de extra agenten, met de inverse methode altijd meer extra agenten ingeroosterd zullen worden dan bij het hanteren van de lineaire methode (in het voorbeeld 39 om 43 agenten).

6.4 Vergelijking

Wat is nu het verschil tussen de twee methoden? Om dit duidelijk te maken, maken we eerst een grafiek waarin we voor beide methoden het aantal in te roosteren agenten uitzetten tegen het percentage shrinkage (uitgaande van een basis staf van dertig agenten):



Figuur 6.2

Zoals eerder vermeld zien we dat er met de inverse methode altijd meer agenten ingeroosterd zullen worden dan met de lineaire methode. Zeker naarmate het percentage shrinkage hoger is, is het verschil erg groot. Uit kostenoverwegingen zou dus waarschijnlijk voor de lineaire methode gekozen worden, dit hoeft echter niet persé tot bevredigende resultaten te leiden. Wat er bij deze methode bijvoorbeeld over het hoofd wordt gezien, is dat een extra agent ook naar het toilet gaat, met collega's overlegt enz. en dus weer voor extra shrinkage zorgt. Met deze extra shrinkage wordt niets gedaan en dus zou het kunnen dat ondanks de extra agenten, het gewenste servicelevel toch niet gehaald wordt.

Bij de inverse methode hebben we wat dat betreft meer zekerheid, toch heeft ook deze methode zijn nadelen. Zo blijkt het niet erg realistisch te zijn om te veronderstellen dat een extra agent hetzelfde percentage shrinkage heeft als een basis agent [2]. De extra agent zal uiteraard zijn pauzes moeten houden maar zal wellicht minder hoeven te vergaderen dan de basis agent, ook zal de kans dat hij ziek is kleiner zijn omdat zijn werktijden op korte termijn zijn geregeld. Daarom zal de inverse methode over het algemeen leiden tot een te hoog aantal agenten en dus onderproductie en te hoge personeelskosten.

Al met al is er dus niet duidelijk te zeggen of één van de twee methoden echt beter is, zeker omdat dit ook per call center verschilt. Zo blijkt het bijvoorbeeld dat voor call centers waar slechts enkele weken vooruit wordt gepland en roosters regelmatig worden aangepast de lineaire methode de werkelijkheid het best benadert. In call centers waar wordt gewerkt met een rooster dat gedurende een grote periode (bijna) hetzelfde is, blijkt de inverse methode weer betere resultaten te geven.

6.5 Een alternatieve methode

Omdat over het algemeen de werkelijke shrinkage tussen het resultaat van de inverse en de lineaire methode zal liggen, heb ik zelf nagedacht over een alternatieve methode:
stel dat iedere extra groep agenten een percentage shrinkage heeft die een factor α keer zo klein is als het percentage van de groep agenten wiens 'shrinkage-gaten' zij moeten opvullen (met $\alpha \geq 1$). Als we bijvoorbeeld $\alpha = 2$, een shrinkage factor van 30% en een basis staf van dertig agenten nemen, krijgen we:

- $\rightsquigarrow$ 1^e groep: 30 agenten $\Rightarrow \frac{30 \cdot 0.3}{2^0} = 9$ extra agenten
- $\rightsquigarrow$ 2^e groep: $\frac{30 \cdot 0.3}{2^0} = 9$ agenten $\Rightarrow \frac{30 \cdot 0.3}{2^0} \cdot \frac{0.3}{2^1} = \frac{30 \cdot 0.3^2}{2^0 \cdot 2^1} = 1.35$ extra agenten
- $\rightsquigarrow$ .....
- $\rightsquigarrow$ n^e groep: $\frac{30 \cdot 0.3^{n-1}}{\prod_{i=0}^{n-2} 2^i}$ agenten $\Rightarrow \frac{30 \cdot 0.3^{n-1}}{\prod_{i=0}^{n-2} 2^i} \cdot \frac{0.3}{2^{n-1}} = \frac{30 \cdot 0.3^n}{\prod_{i=0}^{n-1} 2^i}$ extra agenten

De totale hoeveelheid extra in te roosteren agenten kan nu bepaald worden door het aantal groepen naar oneindig te laten gaan en de som van alle aantallen extra agenten te nemen, ofwel (in het voorbeeld):

$$\textit{extra agenten} = \sum_{n=1}^{\infty} \frac{30 \cdot 0.3^n}{\prod_{i=0}^{n-1} 2^i}$$

of algemeen:

$$\textit{extra agenten} = \sum_{n=1}^{\infty} \frac{\textit{basis staf} \cdot (\textit{\% shrinkage})^n}{\alpha^{n(n-1)/2}}$$

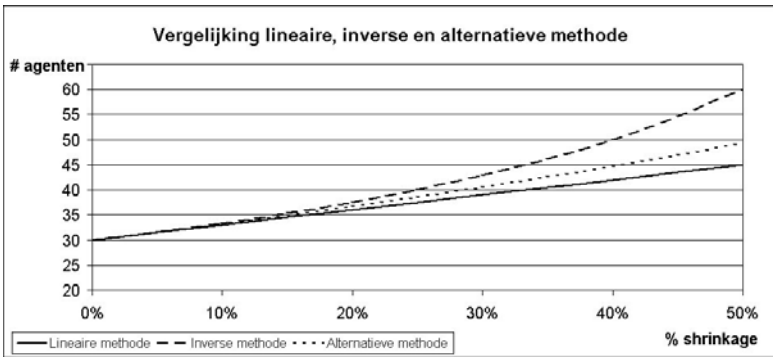
Hierbij maken we gebruik van het feit dat $\prod_{i=1}^{n} \alpha^i = \alpha^{\sum_{i=1}^{n} i}$ en dat $\sum_{i=1}^{n} i = n(n+1)/2$.

Uiteindelijk vinden we dus de volgende formule:

$$\textit{benodigde staf} = \sum_{n=0}^{\infty} \frac{\textit{basis staf} \cdot (\textit{\% shrinkage})^n}{\alpha^{n(n-1)/2}}$$

Merk op dat als we hierin $\alpha = \infty$ invullen (shrinkage tweede groep is 0%) alle termen behalve de eerste twee ($n = 0$ en $n = 1$) gelijk aan nul worden waardoor we precies de formule van de lineaire methode overhouden.
Als we $\alpha = 1$ invullen (shrinkage eerste groep = shrinkage tweede groep = shrinkage derde groep = ...) dan valt de noemer weg en wat we overhouden is precies een meetkundige reeks die de formule voor de inverse methode oplevert.

Voor $1 < \alpha < \infty$ kunnen we nu vrij gemakkelijk de totale shrinkage en dus het aantal benodigde agenten uitrekenen door voor n een groot getal te nemen. In de praktijk blijkt de shrinkage van een afzonderlijke groep al vrij snel naar nul te gaan: als we bijvoorbeeld een basis staf van 250 agenten, een percentage shrinkage van 30% en een α van 2 nemen, dan is de shrinkage van de vijfde groep al nagenoeg gelijk aan nul. Voor $\alpha = 2$ en de overige waarden gelijk aan die van figuur 6.2, is op de volgende pagina de grafiek getekend:



Figuur 6.3

We zien nu dat inderdaad voor ieder percentage shrinkage het aantal benodigde agenten tussen de waarde van de lineaire en die van de inverse methode ligt. Afhankelijk van het soort call center hoeft nu alleen nog maar de waarde van α geschat te worden. Dit kan gedaan worden aan de hand van historische data, bijvoorbeeld met behulp van de meest aannemelijke schatter die besproken is in paragraaf 5.4.

7 Een foute voorspelling, wat nu?

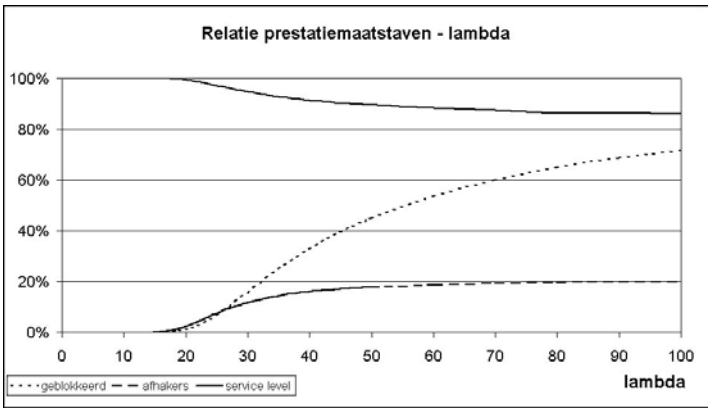
Als we de voorspelling hebben gemaakt, kan het natuurlijk voorkomen dat we er toch naast zitten. Bij een foute voorspelling zijn er of te veel of te weinig agenten aanwezig, wat zal leiden tot een te lage productiviteit dan wel te slechte prestaties richting de klanten. Dit laat zich het best uitleggen aan de hand van de volgende twee voorbeelden, waarbij we ervan uitgaan dat we een fout hebben gemaakt bij het schatten van de aankomstrate. Hetzelfde zou natuurlijk ook gedaan kunnen worden voor de servicerate maar omdat deze makkelijker te schatten is, zal deze situatie zich minder snel voordoen.

Een andere mogelijke fout die we kunnen maken, is een fout geschat percentage shrinkage. Ook dit zal leiden tot te veel of te weinig agenten en dus zien we hier dezelfde consequenties als bij een fout geschatte aankomstrate.

7.1 Slechte service

Stel dat we een call center hebben met een geschat aantal van twintig aankomsten per minuut en een gemiddelde servicetijd van anderhalve minuut. Voor het gemak houden we even geen rekening met shrinkage. Met behulp van een Erlang C calculator kunnen we uitrekenen dat we 34 agenten nodig hebben om een 80/20 servicelevel te behalen. De gemiddelde wachttijd is dan 8.47 seconden. Als nu blijkt dat er in werkelijkheid niet 20 maar 21 klanten per minuut bellen, zien we we dat het servicelevel daalt naar 67,58% en de gemiddelde wachttijd oploopt tot 20.34 seconden. Volgens de Erlang C formule zal een foute voorspelling dus al snel leiden tot een zeer slecht servicelevel.

In werkelijkheid hebben we ook te maken met het blokkeren en afhaken van klanten. Stel dat we veertig lijnen hebben en kijken wat er gebeurt met de verschillende prestatiemaatstaven als we de aankomstrate verkeerd schatten. Dit is weergegeven in figuur 7.1. Hierbij gaan we er weer van uit dat we 34 agenten hebben en dat klanten gemiddeld na één minuut afhaken.



Figuur 7.1

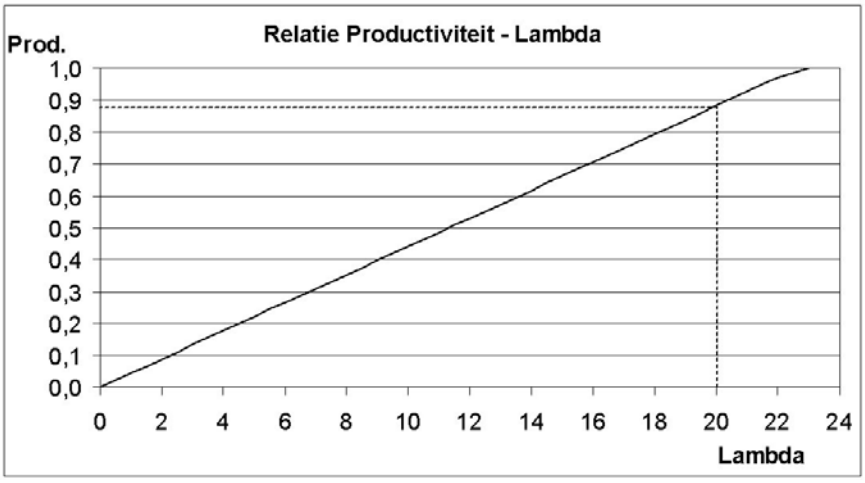
Hierin is duidelijk te zien dat een fout in de voorspelling in werkelijkheid geen dramatische gevolgen heeft voor het servicelevel, dit lijkt namelijk altijd boven de 80% te zullen blijven. Wat we echter wel zien is dat het aantal klanten dat geblokkeerd zal worden sterk toeneemt naarmate de gemaakte fout groter is. Het aantal klanten dat afhaakt lijkt in dit geval richting de 20% te gaan.

Wel zien we dat er eigenlijk niet zoveel aan de hand is als er een kleine fout wordt gemaakt; zowel het percentage geblokkeerde als afgehaakte klanten is lager dan 5% en het servicelevel van de overige klanten is ruim 98% als er in werkelijkheid bijvoorbeeld 23 klanten per minuut blijken te bellen in plaats van de geschatte 20.

Door het aantal lijnen te variëren kan nog naar een optimale balans gezocht worden tussen de drie gemeten prestatiemaatstaven.

7.2 Te lage productiviteit

Stel nu dat (in hetzelfde voorbeeld) er in werkelijkheid niet 20 maar 19 aankomsten per minuut zijn. Het servicelevel stijgt dan weliswaar naar 93.03% (blokkeren en afhaken van klanten laten we even buiten beschouwing), als we echter de productiviteit ($= \lambda\mu/s$) bekijken, zien we dat deze zakt van 88,24% naar 83,82%. Een fout in de voorspelling kan dus ook leiden tot een lagere productiviteit. Dit is ook te zien in de volgende grafiek, waarin de productiviteit is uitgezet tegen de aankomstrate (met dezelfde waarden voor μ en s als bij figuur 7.1):



Figuur 7.2

In het volgende hoofdstuk zal ik nog een manier bespreken om ervoor te zorgen dat de productiviteit hoog blijft, ook al zijn er eigenlijk teveel agenten aanwezig.

8 Risicobeheersing

Zoals we zojuist gezien hebben, kan een foute voorspelling van het aantal in te roosteren agenten vrij makkelijk leiden tot slechte resultaten. Om deze negatieve effecten tegen te gaan, moeten we ervoor zorgen dat ons call center robuuster wordt. Dat wil zeggen dat het minder gevoelig is voor veranderingen en er genoeg tijd is om op veranderingen in te spelen [10]. Dit kan op de volgende manieren.

8.1 Flexibiliteit

Hierbij doelen we op het flexibel kunnen inzetten van personeel bijvoorbeeld door flexibele werktijden. Het nadeel hiervan is dat flexibele werknemers duurder zijn dan werknemers met vaste werktijden, daarom moet de juiste balans gevonden worden tussen vaste en flexibele werknemers. Als je er bijvoorbeeld van uit gaat dat $\lambda \in [19, 22]$ en $\mu = 1.5$, dan zijn er in de rustigste situatie 32 agenten nodig. Als je verder uitgaat van een percentage shrinkage tussen de 25% en 30% dan heb je bij 25% shrinkage dus 40 agenten nodig voor een 80/20 servicelevel (lineaire methode). Mocht nu alles tegenzitten, dus $\lambda = 22$ en het percentage shrinkage is 30%, dan heb je 49 agenten nodig om alsnog het gewenste servicelevel te halen. In dat geval zou je dus 40 vaste agenten kunnen inroosteren en 9 flexibele agenten achter de hand houden.

Figuur 8.1

8.2 Cross training

Door agenten te leren om meerdere taken uit te voeren, kan er ook voor gezorgd worden dat het call center robuuster wordt. Als het in een bepaalde groep rustig is, kan een agent uit die groep bijspringen in een andere groep waar het wel druk is. In het voorbeeld van de multinational kan het dus makkelijk zijn om meertalige agenten te hebben. Een agent die zowel Nederlands als Engels spreekt kan dan op bepaalde momenten Nederlandse klanten te woord staan en op andere momenten bijvoorbeeld Amerikaanse.
Het bijspringen van stafmedewerkers of managers in geval van drukte valt ook onder cross training.

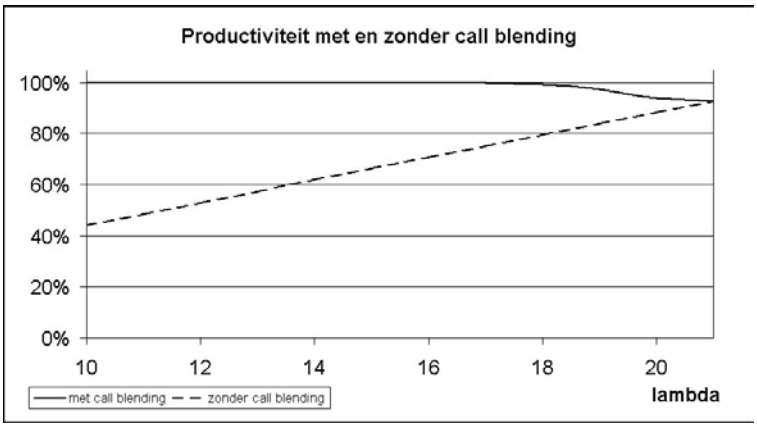
8.3 Meerdere kanalen

Een andere manier om het call center robuust te maken is om meerdere kanalen te gebruiken. Omdat het beantwoorden van emails een ander servicelevel vereist dan het beantwoorden van telefoongesprekken (meestal binnen één dag) kunnen deze afgehandeld worden op de rustige momenten van de dag. Op de drukke momenten kunnen alle agenten zich dan bezighouden met het beantwoorden van binnenkomende telefoongesprekken.

8.4 Call blending

In het verlengde van het gebruik van meerdere kanalen ligt ten slotte nog het zogenaamde call blending om het call center robuuster te maken. Bij call blending worden er ook meerdere kanalen gebruikt, maar dan tegelijkertijd. Als op een bepaald moment een aantal agenten geen gesprek voert, krijgt een deel van hen (automatisch) een uitgaand gesprek of email om af te handelen. Met $\lambda = 19$ en $\mu = 1.5$ zagen we al eerder dat er 40 agenten nodig zijn (inclusief shrinkage). Stel nu dat er op een dag helemaal geen zieken zijn, dan zijn er dus eigenlijk teveel agenten aanwezig. We kunnen dan echter naast de binnenkomende gesprekken ook nog vijf emails of uitgaande gesprekken per minuut afhandelen en toch nog een 80/20 servicelevel halen. De productiviteit is nu 100% tegenover 79.2% zonder call blending. We zien dus dat de productiviteit toch nog hoog is, ondanks een te hoog ingeschat aantal agenten.

Bovenstaand voorbeeld is uitgerekend met behulp van de calculator die te vinden is op www.math.vu.nl/∼koole/ccmath/blending.php. In figuur 8.2 is de productiviteit bij gebruik van call blending uitgezet naast de productiviteit zonder call blending. Hierbij is vereist dat er altijd een 80/20 servicelevel gescoord moet worden. Voor $\lambda \geq 22$ is dit niet meer mogelijk.



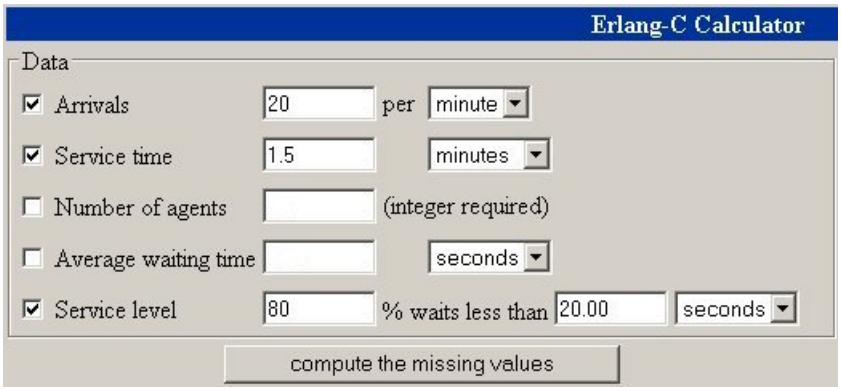
Figuur 8.2

We zien dat er met call blending bijna altijd een productiviteit van rond de 100% gescoord wordt. Hierdoor is dit een krachtig wapen om het call center robuust te maken, er kan immers altijd gezorgd worden voor een hoge productiviteit, ook al zijn er eigenlijk teveel agenten ingeroosterd. Uiteraard moet er wel voor gezorgd worden dat er altijd voldoende uitgaande gesprekken of emails zijn om af te handelen.

Merk op dat voor grote waarden van λ de productiviteit met gebruik van call blending gelijk is aan de productiviteit zonder gebruik hiervan. Dit komt doordat er voor $\lambda \geq 21$ geen uigaande gesprekken meer gevoerd worden, alle agenten zijn dan namelijk nodig om binnen komende gesprekken af te handelen.

9 Het maken van de planning

Als eindelijk de aankomst- en servicerates zijn geschat en we hebben nagedacht over hoe we om zullen gaan met shrinkage, kunnen we de planning gaan maken. Om een basisoplossing te vinden gebruiken we een Erlang calculator. Vanwege de eenvoud van de Erlang C formule, heb ik deze gebruikt in de voorbeelden.



Figuur 9.1

Dit is de Erlang calculator die gevonden kan worden via de in paragraaf 5.3 genoemde link. In de eerste twee vakjes vullen we de schattingen in voor de aankomst- en servicerate (bijvoorbeeld $\lambda = 20$ en $\mu = 1.5$) en vinken het keuzevakje aan. Omdat we in het aantal benodigde agenten en de gemiddelde wachttijd geïnteresseerd zijn, laten we hier de keuze- en tekstvakjes leeg. Tot slot vullen we het gewenste servicelevel in, ik heb hier gekozen voor het standaard 80/20 servicelevel. Als we nu op de bereken-knop drukken, vinden we dat er minimaal 34 agenten nodig zijn om het gewenste servicelevel te bereiken.

Stel nu dat we ons wat meer willen indekken tegen het risico van een fout voorspelde aankomstrate (over de servicerate hoeven we ons zoals eerder vermeld minder zorgen te maken). Op basis van historische data hebben we als 95% betrouwbaarheidsinterval gevonden dat $\lambda \in [19, 22]$. Met een vast aantal agenten van 34 betekent dit dat de 95% betrouwbaarheidsintervallen voor het servicelevel (SL) en de gemiddelde wachttijd (AWT) worden gegeven door $SL \in [35.37\%, 93.03\%]$ respectievelijk $AWT \in [3.87 \text{ sec.}, 72.65 \text{ sec.}]$.

We kunnen nu dus twee dingen doen:

- ⇝ Omdat er een grote kans bestaat dat we het gewenste servicelevel niet halen, kunnen we in dit geval uitgaan van het zogenaamde *worst-case-scenario* dat $\lambda = 22$. We hebben dan 37 agenten nodig om een 80/20 servicelevel te halen, nu hebben we echter te maken met het risico dat de productiviteit te laag wordt (in het geval dat λ toch een stuk lager is dan 22).
  In dit geval zou de productiviteit opgekrikt kunnen worden door te zorgen voor voldoende uitgaande gesprekken die afgehandeld kunnen worden door het overschot aan agenten.

- ⇝ We gaan uit van de meest rustige situatie, dus $\lambda = 19$. In dat geval hebben we 32 agenten nodig om het gewenste servicelevel te halen. Omdat we weten dat de aankomstrate kan stijgen tot $\lambda = 22$ en dat dit zal leiden tot 37 benodigde agenten, houden we vijf flexibele agenten achter de hand die we inschakelen als blijkt dat de aankomstrate inderdaad hoger is.

Afhankelijk van de kosten van flexibele en vaste agenten en de beschikbare hoeveelheid alternatieve werkzaamheden moet er een keuze gemaakt worden. Uiteraard is het ook mogelijk om te kiezen voor een combinatie van de twee mogelijkheden, bijvoorbeeld 34 vaste agenten inroosteren en 3 flexibele agenten achter de hand houden.

Omdat we nog geen rekening hebben gehouden met het optreden van shrinkage, moeten we nog een aantal extra agenten inroosteren. Hierbij gebruiken we de formule die ik in paragraaf 6.5 heb besproken. Stel dat we op basis van historische data een percentage shrinkage van 28% hebben geschat en dat we er van uit gaan dat de shrinkage voor iedere volgende groep agenten gehalveerd wordt (dus $\alpha = 2$), afhankelijk van de gekozen methode (worst case of flexibiliteit, zie hierboven) besluiten we om 49 respectievelijk 43 agenten in te roosteren.

10 Conclusies

Omdat er bij het voorspellen van met name de aankomstrate veel onzekerheid om de hoek komt kijken, is het belangrijk dat hier veel aandacht aan besteed wordt. Met behulp van de in paragraaf 5.4 besproken Poisson mixtures methode kan een betrouwbaarheidsinterval worden opgesteld voor de aankomstrate. Voor de servicerate kan het gemiddelde worden gebruikt omdat deze minder afhankelijk is van externe factoren en omdat er meestal meer relevante data over beschikbaar is.

Aan de hand van de schattingen voor de aankomst- en servicerate en het gewenste servicelevel kan met behulp van de Erlang C formule het aantal benodigde agenten bepaald worden. Deze formules zijn vaak geïmplementeerd in beslissingsondersteunende systemen, bijvoorbeeld op
www.math.vu.nl/∼koole/ccmath/ErlangC.php.

Daarnaast moet er ook rekening worden gehouden met niet-productieve uren van de agenten, zoals bijvoorbeeld ziekteverzuim of werkoverleg. Deze verloren tijd wordt ook wel shrinkage genoemd. Ik denk dat het niet verstandig is om hiervoor, zoals gebruikelijk is, een vast percentage van rond de 30% te kiezen. In veel gevallen zal er namelijk al ruim van te voren bekend zijn wanneer en bij welke agenten shrinkage zal optreden, dit is bijvoorbeeld het geval bij vergaderingen. Daarom is het beter om agenten waarvan bekend is dat ze op een bepaald dagdeel een vergadering hebben gewoon niet in te roosteren voor hun normale werkzaamheden. Overigens kan ik me haast niet voorstellen dat dit niet overal gebeurt, maar in alle literatuur die ik heb onderzocht worden vergaderingen, pauzes enz. toch tot shrinkage gerekend...

Uiteraard moet er ook rekening gehouden worden met ziekteverzuim van agenten. De manier waarop dit meestal gebeurt is dat er een percentage van rond de 10% gehanteerd wordt. Ik denk dat dit percentage wat aan de hoge kant is; gemiddeld zal het misschien wel 10% zijn, maar ik vermoed dat de variantie van het ziekteverzuim vrij groot zal zijn. Daarom is het volgens mij beter om het benodigde aantal agenten te verhogen met een laag percentage van bijvoorbeeld 5%. Mochten er op een bepaalde dag toch meer agenten ziek zijn, dan kan dit worden opgevangen door het inzetten van flexibele agenten zoals bijvoorbeeld studenten. Verder kan het denk ik geen kwaad om onderzoek te doen naar de kansverdeling van het ziekteverzuim. Helaas heb ik dit niet kunnen doen omdat er geen data beschikbaar was.

Het belangrijkste aspect is echter dat ervoor gezorgd moet worden dat het call center zodanig is ingericht dat het snel kan reageren op gewijzigde omstandigheden. Door de ondergrens van het gevonden betrouwbaarheidsinterval van de aankomstrate te gebruiken voor het bepalen van het aantal vaste agenten dat ingeroosterd moet worden en het verschil tussen de bovengrens en de ondergens voor het aantal flexibele agenten, kan hiervoor gezorgd worden.

Ook door meerdere kanalen en call blending te gebruiken kan er goed gereageerd worden op onverwacht rustige momenten (paragraaf 8.3 en 8.4). Een aantal agenten kan zich dan bezighouden met uitgaande gesprekken terwijl de rest beschikbaar blijft voor binnenkomende gesprekken. Hierdoor blijft de productiviteit hoog, ondanks het feit dat er eigenlijk teveel agenten aanwezig zijn.

Ten slotte kan er nog voor gekozen worden om het aantal lijnen te beperken. Hierdoor zullen er klanten geblokkeerd worden als alle lijnen bezet zijn waardoor de wachtrij nooit te lang zal worden. Het servicelevel van de klanten die er wel in geslaagd zijn om het systeem binnen te komen, zal hierdoor over het algemeen wel goed zijn (zie figuur 7.1). Het nadeel hier is echter dat de klanten die geblokkeerd worden dit waarschijnlijk als zeer frustrerend zullen ervaren en wellicht een andere keer terugbellen. Daarom is dit denk ik geen goede methode; eigenlijk is er slechts sprake van uitstel van executie in plaats van dat het probleem echt opgelost wordt.

11 Case studie

In dit hoofdstuk wil ik de in het werkstuk gebruikte methoden toepassen op 'real world data'. De case studie zal opgebouwd zijn uit twee onderdelen. In het eerste deel zal gekeken worden naar het bepalen van het aantal benodigde agenten. Hiervoor heb ik de data van Avi Mandelbaum gebruikt. Dit is data van een anonieme bank in Israel en is vrij beschikbaar op http://iew3.technion.ac.il/serveng/. In het tweede deel wil ik de verschillende methoden hoe met shrinkage omgegaan kan worden met elkaar vergelijken. Hierbij zal ik vooral het ziekteverzuim bekijken. Omdat hiervoor geen data beschikbaar was, heb ik een fictief voorbeeld uitgewerkt.

11.1 Voorspellen aankomst- en servicerate

De originele data bestaat uit veel variabelen, hierin is onder andere het type gesprek opgenomen. Voor de analyse heb ik alleen de reguliere gesprekken (aangeduid met *type=PS*) gebruikt. Er zijn ook enkele virtuele gesprekken in de database opgenomen (aangeduid met *outcome=PHANTOM*), deze heb ik eerst verwijderd. Ik heb alleen de data van doordeweekse dagen gebruikt omdat de aankomsttijden van de gesprekken in het weekend waarschijnlijk een andere verdeling zullen hebben dan doordeweeks. Omdat het call center op vrijdag om 14:00 uur sluit, heb ik alleen de intervallen tussen 7:00 en 14:00 bekeken.
In de gegeven voorbeelden zal ik gebruik maken van het interval 10:00-10:30 in de maand februari. Om de aankomstrate te schatten heb ik een programma geschreven dat het aantal aankomsten per periode per dag telt, hiermee heb ik tabellen van de volgende vorm gemaakt:

	Dag 1	Dag 2	Dag 3	...
07:00-07:30	18	14	10	...
07:30-08:00	20	27	27	...
08:00-08:30	50	35	34	...
...	...	...	...	⋱

Tabel 11.1

De complete tabel voor de maand februari is te vinden in appendix B. Voor het interval 10:00-10:30 vinden we gemiddeld 38.65 aankomsten per half uur. Als we het 95% betrouwbaarheidsinterval, gedefineerd als in paragraaf 5.1 bepalen, vinden we dat $\lambda \in [35.93, 41.37]$. Als we echter naar de data kijken (appendix B), zien we dat er al negen dagen zijn dat de werkelijke λ groter is dan de bovengrens. Op deze dagen zal er dus hoe dan ook een tekort aan agenten zijn. Blijkbaar hebben we bij de gebruikte methode foute aannames gemaakt. Om te testen of de aanname dat aankomsttijden exponentieel(λ) verdeeld zijn correct is, maak ik gebruik van de eigenschap dat het aantal aankomsten in een bepaalde periode Δt dan een Poisson($\lambda \Delta t$) verdeling moet hebben [14].

Zoals eerder vermeld moeten het gemiddelde en de variantie dan ongeveer aan elkaar gelijk zijn, we vinden hier 38.65 respectievelijk 102.66. Dit is bij lange na niet aan elkaar gelijk.

Als we de Neyman-Scott test (een toets waarmee we kunnen testen of een bepaalde dataset een Poisson verdeling heeft) toepassen, vinden we dan ook een p-waarde van nul en dus verwerpen we de nulhypothese dat het aantal aankomsten een Poisson verdeling heeft [8].

Kennelijk was de aanname van exponentieel verdeelde aankomsttijden niet goed. Daarom zal ik met behulp van de in paragraaf 5.4 besproken Poisson mixtures methode kijken of dit een betrouwbaardere schatting geeft voor de aankomstrate. Hiervoor schatten we eerst de parameters $\hat{r}$ en $\hat{\lambda}$, we vinden $\hat{r} = 10.6$ en $\hat{s} = 0.4$. Nu kunnen we de gammakwantielen (met $\alpha = 0.25$) bepalen en vinden voor het aantal aankomsten per half uur het interval $[13.79, 47.37]$.

Zoals gezegd kunnen we voor de servicetijd de puntschatter gebruiken, we vinden hier een gemiddelde servicetijd van 186.18 seconden.

Met behulp van de Erlang formule kunnen we nu het aantal benodigde agenten uitrekenen. Voor $\lambda = 13.79$ vinden we drie benodigde agenten en voor $\lambda = 47.37$ zijn dit er acht.
Dit betekent dat het waarschijnlijk het verstandigst is om drie vaste agenten in te roosteren en vijf flexibele agenten achter de hand te houden.

Naarmate er voldoende uitgaande gesprekken of emails zijn af te handelen, kan er ook voor gekozen worden om meer vaste agenten in te roosteren, bijvoorbeeld zes vaste agenten en twee flexibele.

11.2 Shrinkage

Ik heb hier helaas geen onderzoek naar kunnen doen omdat er geen data beschikbaar was. Ik denk dat het voornaamste punt hier is om achter de kansverdeling van het aantal (of percentage) zieke agenten te komen. Hiervoor is bijvoorbeeld een tabel van de volgende soort nodig:

	Aantal ingeroosterd	Aantal zieken
Dag 1	34	3
Dag 2	36	3
Dag 3	33	5
Dag 4	35	1
...	...	...

Tabel 11.2

Er zijn dan verschillende statistische methoden om de gevraagde verdeling te bepalen [5]. Omdat er geen echte data beschikbaar is, zal ik hieronder een fictief voorbeeld uitwerken. Hierbij ga ik ervan uit dat er voor een bepaald interval iedere dag 37 agenten nodig zijn. Omdat het ziekteverzuim gemiddeld ongeveer 10% is, worden er steeds 41 agenten ingeroosterd (dus vier agenten extra).

Omdat ik denk dat het aantal zieken op de meeste dagen laag zal zijn en op sommige dagen hoog (*"het heerst!"*), heb ik gekozen voor een lognormale verdeling met een gemiddelde van 3.7 en standaardafwijking 10. Met behulp van Crystal Ball heb ik 10000 keer een periode van 50 dagen gesimuleerd. De frequentiegrafieken zijn weergegeven in bijlage C. We zien nu dat er, ondanks dat we vier agenten extra ingeroosterd hebben om ziektegevallen op te vangen, dat er toch ongeveer 21.68% van de dagen (10.84 van de 50) een tekort aan agenten zal zijn. Op dagen dat er een tekort is, is het gemiddelde tekort 8.87 agenten, wat betekent dat er op deze dagen gemiddeld negen flexibele agenten nodig zijn. Aan de andere kant zijn er op dagen met een overschot (39.16 van de 50) gemiddeld 2.82 agenten teveel, wat de productiviteit niet ten goede komt (tenzij er gebruik wordt gemaakt van call blending).

Als we nu besluiten om het anders te doen, namelijk om helemaal geen extra agenten in te roosteren en alleen gebruik te maken van flexibele agenten, dan zien we dat er iedere dag een tekort aan agenten is. Dit zal daarom waarschijnlijk niet tot bevredigende resultaten leiden. Daarom besluiten we om één of twee extra agenten in te roosteren in plaats van vier, we vinden dan de resultaten die gegeven zijn in tabel 11.3.

Hierin staat in de tweede kolom het verwachte overschot aan agenten, ofwel $\mathbb{E}$[tekort]. In de derde kolom zien we het verwachte tekort op dagen dat er een tekort is (dus $\mathbb{E}$[tekort|tekort]). In de vierde kolom zien we dit in geval van een overschot.

	Gem. aantal dagen tekort	Gem. tekort	Gem. tekort bij een tekort	Gem. overschot bij een overschot
één extra	56.70%	2.71	5.21	0.55
twee extra	37.94%	1.71	6.58	1.25

Tabel 11.3

We zien dus er, als we slechts één agent extra inroosteren, op de meeste dagen (56.70%) nog een tekort is. Op deze dagen zouden er vijf á zes flexibele agenten nodig zijn om dit tekort op te vangen (vierde kolom). Op de dagen dat er geen tekort is, zien we dat er nog maar een klein overschot is (gemiddeld 0.55 agenten).
Iets beter is het volgens mij om twee agenten extra in te roosteren. Er is dan op ongeveer 37.94% van de dagen een tekort aan agenten en op deze dagen zijn er gemiddeld zes á zeven flexibele agenten nodig. Op dagen met een overschot zijn er maar weinig agenten teveel (gemiddeld 1.25). Het gemiddelde tekort is 1.71 agenten, wat ook logisch is omdat er gemiddeld 3.7 agenten per dag ziek zijn en er slechts twee extra ingeroosterd worden. Op deze manier wordt er voor gezorgd dat er zelden een (groot) overschot aan agenten is, waardoor de productiviteit altijd hoog zal zijn.

Overigens is de keuze voor de hoeveelheid in te roosteren extra agenten grotendeels afhankelijk van de hoeveelheid werk die beschikbaar is voor een eventueel overschot aan agenten. In paragraaf 8.4 hebben we immers gezien dat een overschot aan agenten helemaal niet zo erg hoeft te zijn, als er maar voldoende ander werk voor hen is, bijvoorbeeld uitgaande gesprekken voeren.

Als er niet genoeg werk te doen is voor agenten in geval van een overschot, bijvoorbeeld in een call center met alleen binnenkomende gesprekken, dan denk ik echter dat het verstandiger is om weinig extra agenten in te roosteren voor ziektegevallen en zoveel mogelijk te werken met flexibele agenten...

Merk hierbij nog wel op dat het waarschijnlijk niet juist is om te veronderstellen dat het ziekteverzuim lognormaal verdeeld is. Dit is enkel een door mij gekozen verdeling, gekozen op basis van mijn intuïtie en zonder enig verdelingsonderzoek. Bovendien is de lognormale verdeling een absoluut continue verdeling terwijl de werkelijke verdeling van het aantal zieke agenten natuurlijk discreet zal zijn.
Om een realistisch vergelijk te krijgen zou met behulp van historische data de ware verdeling geschat moeten worden waarna een analyse zoals hierboven uitgevoerd kan worden.

Appendix A: gebruikte kansverdelingen

De Poisson verdeling .
Een stochastische variabele X heeft een Poisson verdeling met parameter λ als deze discreet verdeeld is en voor de kansdichtheid $p(k)$ in het punt $k \in \mathbb{Z}_+$ geldt dat:

$$p(k) = \mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Voor de verwachting en variantie geldt dat $\mathbb{E}X = \mathrm{Var}X = \lambda$.
De Poisson verdeling wordt vaak gebruikt als kansverdeling voor het aantal incidenten in een bepaalde periode, zoals in ons geval het aantal binnenkomende telefoongesprekken in Δt tijdseenheden.

De geometrische verdeling .
Een discreet verdeelde stochastische variabele X heeft een geometrische verdeling met parameter $p \in (0, 1)$ als geldt dat:

$$p(k) = \mathbb{P}(X = k) = p(1-p)^{k-1}$$

De interpretatie van de geometrische verdeling is het aantal onafhankelijke Bernoulli proeven met succeskans p, nodig voor het behalen van een succes.
De staartkans $\mathbb{P}(X > k)$ kan bepaald worden door:

$$\begin{aligned} \mathbb{P}(X > k) &= 1 - \mathbb{P}(X \leq k) \\ &= \textstyle\sum_{n=1}^{k} p(1-p)^{n-1} \\ &= (1-p)^k \end{aligned}$$

waarbij we gebruik maken van het feit dat we te maken hebben met een meetkundige reeks.
Als verwachting en variantie vinden we dat $\mathbb{E}X = 1/\mathrm{p}$ en $\mathrm{Var}X = (1-p)/p^2$.

De geometrische verdeling heeft een nuttige eigenschap, hij is namelijk geheugenvrij in discrete zin, d.w.z. dat $\mathbb{P}(X > k + s | X > k) = \mathbb{P}(X > s)$:

$$\begin{aligned} \mathbb{P}(X > k + s | X > k) &= \mathbb{P}(X > k + s \cap X > s)/\mathbb{P}(X > k) \\ &= \mathbb{P}(X > k + s)/\mathbb{P}(X > k) \\ &= (1-p)^{k+s}/(1-p)^k \\ &= (1-p)^s \\ &= \mathbb{P}(X > s) \end{aligned}$$

De negatief binomiale verdeling .
We zeggen dat een stochastische variabele X een negatief binomiale verdeling heeft als deze discreet verdeeld is en er voor $p(k)$ geldt dat:

$$p(k) = \mathbb{P}(X = k) = \left(\begin{array}{c} k-1 \\ r-1 \end{array} \right) p^r (1-p)^{k-r}$$

De bijbehorende paramters zijn $r \in \mathbb{N}$ en $p \in (0, 1)$. De interpretatie van de negatief binomiale verdeling is dat het aantal onafhankelijke Bernoulli proeven met succeskans p dat nodig is voor het behalen van r successen negatief binomiaal verdeeld is. Merk op dat als we $r = 1$ invullen, we precies een geometrische verdeling overhouden.
Als verwachting en variantie vinden we $\mathbb{E}X = \mathrm{r/p}$ en $\mathrm{Var}X = r(1-p)/p^2$.

De exponentiële verdeling .
Een stochastische variabele X heeft een exponentiële verdeling met parameter $\lambda > 0$ als deze absoluut continu verdeeld is en de kansdichtheid $f(x)$ gegeven wordt door:

$$f(x) = \lambda e^{-\lambda x}$$

De verdelingsfunctie $F(x)$ wordt dan gegeven door:

$$\begin{array}{rl} F(x) & = \int_0^x \lambda e^{-\lambda x} \\ & = 1 - e^{-\lambda x} \end{array}$$

De verwachting en variantie van een exponentieel(λ) verdeelde variabele X zijn $\mathbb{E}X = 1/\lambda$ respectievelijk $\mathrm{Var}X = 1/\lambda^2$.

Net als de geometrische verdeling is de exponentiële verdeling geheugenvrij (maar dan in absoluut continue zin):

$$\begin{array}{rl} \mathbb{P}(X > s+t | X > s) & = \mathbb{P}(X > s+t \cap X > s)/\mathbb{P}(X > s) \\ & = \mathbb{P}(X > s+t)/\mathbb{P}(X > s) \\ & = e^{-\lambda(s+t)}/e^{-\lambda s} \\ & = e^{-\lambda s} e^{-\lambda t}/e^{-\lambda s} \\ & = e^{-\lambda t} \\ & = \mathbb{P}(X > t) \end{array}$$

Dit houdt in dat als de aankomsttijd exponentieel verdeeld is, er geen rekening mee gehouden hoeft te worden hoe lang er al geen aankomst is geweest; de kans op een aankomst in de komende t tijdseenheden is altijd $\mathbb{P}(X \leq t) = 1 - e^{-\lambda t}$.

De gamma verdeling .
Een (absoluut continue) stochastische variabele X heeft een Gamma verdeling (ook wel Erlang verdeling) met parameters $r \in \mathbb{N}$ en $\lambda > 0$ als de kansdichtheid $f(x)$ gegeven wordt door:

$$f(x) = \frac{\lambda^r}{(r-1)!} x^{r-1} e^{-\lambda x}$$

Voor de verdelingsfunctie $F(x)$ is geen eenvoudige uitdrukking te geven. De verwachting en variantie zijn $\mathbb{E}X = \mathrm{r}/\lambda$ en $\mathrm{Var}X = r/\lambda^2$.

Merk op dat als we $r = 1$ nemen, we een exponentiële verdeling met parameter λ overhouden. We zeggen dan ook wel dat een exponentiële verdeling tot dezelfde locatie-schaal familie behoort als een gamma verdeling.

De normale verdeling .
Een stochastische variabele X is normaal verdeeld met parameters μ en σ^2 als deze absoluut continu is en voor de kansdichtheid $f(x)$ geldt dat:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

De verwachting is gelijk aan $\mathbb{E}X = \mu$ en de variantie is $\mathrm{Var}X = \sigma^2$.
Als we $\mu = 0$ en $\sigma = 1$ invullen dan hebben we te maken met een standaard normale verdeling.

De lognormale verdeling .
Een stochastische variabele X heeft een lognormale verdeling als hij absoluut continu verdeeld is en zijn natuurlijke logaritme $Y = \log(X)$ een normale verdeling heeft. De dichtheidsfunctie wordt (voor $x > 0$) gegeven door:

$$\Phi(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log(x)-\mu}{\sigma}\right)^2}$$

De verwachting is $\mathbb{E}X = \mathrm{e}^{(2\mu+\sigma^2)/2}$.
Voor de variantie vinden we $\mathrm{Var}X = e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2}$.

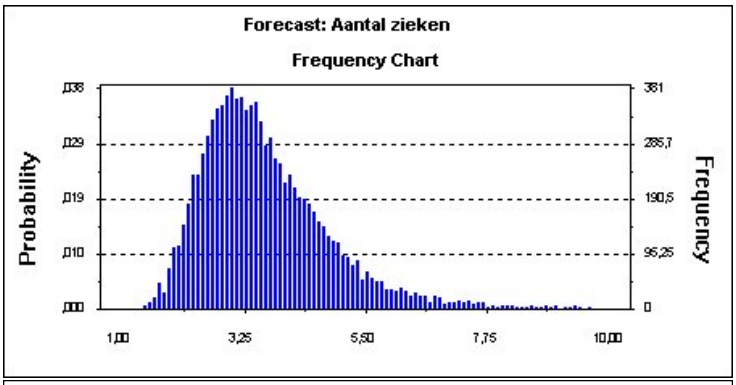
Appendix B: Data februari

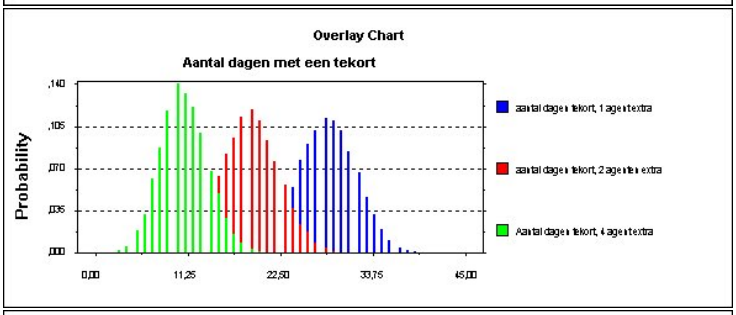
In de volgende tabel staat, voor de maand februari, per dag hoeveel (reguliere) aankomsten er waren in de verschillende tijdsintervallen van een half uur.
De originele data is afkomstig van het call center van een anonieme Israelische bank en is beschikbaar op http://iew3.technion.ac.il/serveng/.

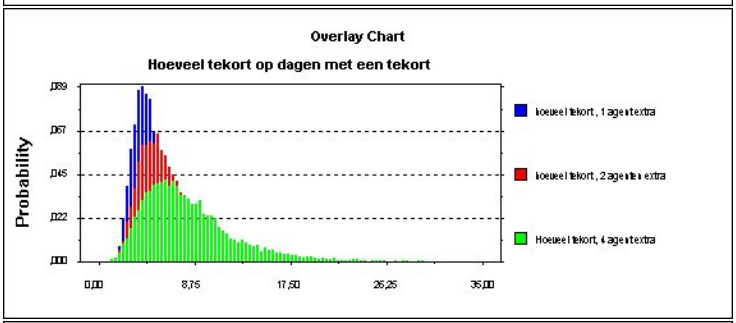
	7:00	7:30	8:00	8:30	9:00	9:30	10:00	10:30	11:00	11:30	12:00	12:30	13:00	13:30
1 feb	18	20	50	29	32	49	39	45	38	54	34	30	31	41
2 feb	14	27	35	42	43	47	58	52	50	36	41	38	37	59
3 feb	10	27	34	30	33	43	45	50	51	40	28	40	374	73
4 feb	9	14	15	28	35	29	49	38	30	30	33	30	25	29
5 feb	6	17	23	25	21	25	32	25	40	34	23	18	19	10
8 feb	6	12	11	15	18	31	36	46	29	28	26	39	8	35
9 feb	8	6	16	18	18	12	25	15	15	37	26	19	15	16
10 feb	20	28	28	35	32	46	45	49	45	42	37	33	29	33
11 feb	14	31	27	44	47	42	43	63	55	46	43	30	41	35
12 feb	4	14	22	28	32	29	34	18	26	27	19	18	17	15
15 feb	12	14	30	28	26	48	37	39	43	36	23	34	34	35
16 feb	17	29	39	46	42	57	52	55	45	49	30	34	43	51
17 feb	13	12	20	29	41	45	41	41	40	41	28	28	46	50
18 feb	10	16	17	35	34	31	43	37	32	34	30	28	36	30
19 feb	6	11	15	20	25	21	16	25	33	22	21	17	19	18
22 feb	24	27	25	35	38	62	47	48	43	44	37	31	36	39
23 feb	16	18	15	30	29	34	43	44	28	41	35	22	30	34
24 feb	9	12	27	20	23	37	36	34	29	24	22	31	33	21
25 feb	15	18	17	24	31	38	39	38	45	24	35	30	32	33
26 feb	8	8	18	32	20	20	21	18	18	32	24	21	12	8

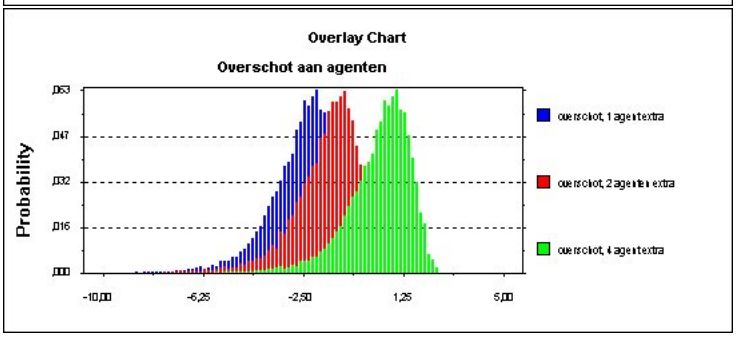
De ontbrekende dagen zijn de weekenden, deze zijn niet meegenomen in het onderzoek. Merk op dat in een kolom steeds het interval **vanaf** het tijdstip bovenaan staat, dus in de tweede kolom het interval 7:00-7:30, in de derde het interval 7:30-8:00 enz.

Appendix C: Frequentie grafieken ziekteverzuim









Referenties

[1] *A. Avramidis, A. Deslauriers & P. L'Ecuyer - Modeling daily arrivals to a telephone call center - Management science, 2004*

[2] *G. Barber - Two ways to think about shrinkage - Call center magazine, 2004*

[3] *B. Cleveland - 12 traits of the best managed call centers - ICMI, 2002*

[4] *P. Dehaan - Call center shrinkage - Connections magazine, 2004*

[5] *M. de Gunst & A. van der Vaart - Statistische data analyse - Collegedictaat*

[6] *K. van Harn & P. Holewijn - Inleiding waarschijnlijkheidsrekening - Collegedictaat*

[7] *G. Koole - Call center mathematics - Collegedictaat*

[8] *G. Koole & G. Jongbloed - Managing uncertainty in call centers using Poisson mixtures - Applied Stochastic Models in Business and Industry, 2001*

[9] *G. Koole - Modeling of business processes - Collegedictaat*

[10] *G. Koole & A. van Moorst - WFM: goede voorspelling noodzakelijk maar niet voldoende - Telecommerce, 2004*

[11] *J. Oosterhoff & A. van der Vaart - Algemene statistiek - Collegedictaat*

[12] *P.Reynolds - A new look at the call center - The call center school, 2002*

[13] *P. Reynolds - The math of call center staffing - The call center school, 2003*

[14] *H. Tijms - A first course in stochastic models - J. Wiley % Sons, 2003*

[15] *H. Tijms - Operationele analyse - Epsilon 2002*

[16] *W. Whitt - Staffing a call center with uncertain arrival rate and absenteeism - Management science, 2004*