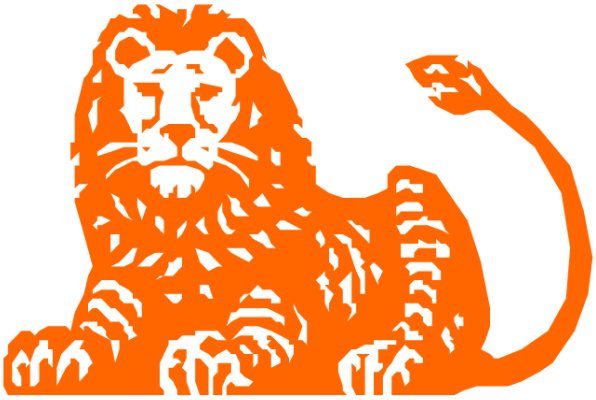
Versneld Evalueren

Een onderzoek naar het voorspellen van respons
voor marketingcampagnes



ING NEDERLAND

Stageverslag BWI
Francesca Armandillo

Niet–vertrouwelijke versie

Maart 2007

Versneld Evalueren

Een onderzoek naar het voorspellen van respons
voor marketingcampagnes

Stageverslag BWI
Francesca Armandillo
Begeleiding: Dr. W. Kowalczyk, Dr. M. de Gunst, Drs. W.B.Tip

Vrije Universiteit Amsterdam
Faculteit der Exacte Wetenschappen
Bedrijfswiskunde & Informatica
De Boelelaan 1081
1081 HV Amsterdam

Stagebedrijf:
ING Nederland
Retail / Customer Intelligence / Research & Modelling
Haarlemmerweg 520
1014 BL Amsterdam

Maart 2007

Voorwoord

Het laatste onderdeel van de studie Bedrijfswiskunde en Informatica is een afstudeerstage. Ik heb mijn stage gelopen bij ING Retail op de afdeling Customer Intelligence/Research & Modelling. Tijdens deze stage heb ik onderzoek gedaan naar het voorspellen van het responsverloop van marketing campagnes op basis van het begin van de responsperiode. Dit onderzoek is voor het product Hypotheken uitgewerkt.

Eerst wil ik graag een aantal mensen bedanken voor het goede verloop en uitstekende begeleiding van de stage. Ten eerste mijn begeleider bij de ING, Wim Tip. Wim heeft mij een grote mate van vrijheid gegeven om de aanpak en uitvoering van de opdracht uit te voeren. Verder wil ik graag mijn begeleiders van de VU, Wojtek Kowalczyk en Mathisca de Gunst, bedanken voor de inzet en hulp gedurende de stage.
Ook wil ik graag de overige collega's van Research & Modelling hartelijk bedanken voor de gezellige sfeer, enthousiasme en openheid voor het stellen van vragen.

Tenslotte bedank ik mijn ouders die het voor mij mogelijk hebben gemaakt deze studie te volgen.

Francesca Armandillo

Amsterdam, maart 2007

Samenvatting

In de wereld van marketing wil men de resultaten van de campagnes zo snel mogelijk gebruiken voor nieuwe campagnes. Bij het product hypotheken is de responsperiode lang. Als vroeg in de actie al geëvalueerd kan worden hoeveel hypotheken de actie gaat opbrengen, en dus hoe effectief een actie is, dan kan redelijk vroeg in de actie worden ingezien of een actie genoeg hypotheken zal gaan opleveren, en dus of deze winstgevend zal zijn of niet.

Als eerste is voor het product hypotheken onderzocht in welke periode de productie van de actie zit. Het doel is namelijk om deze productie zo vroeg mogelijk in de tijd te kunnen voorspellen. Nadat deze productie is gelokaliseerd wordt het onderzoek op vier manieren benaderd. Allereerst is met behulp van *tijdreeks analyse* geprobeerd het responsverloop te modelleren. Een andere benadering is meer op *klant niveau* gericht: *zijn er klanten met bepaalde eigenschappen die altijd vroeg of laat responderen*? Daarna is onderzocht of bepaalde begin dagen goede voorspellers zijn voor de totale respons gebaseerd op gemiddelde responspercentages van historische acties. Als laatste is het responsverloop door een poisson proces gemodelleerd.

Uit dit onderzoek volgt dat het mogelijk is de productie van een actie te voorspellen. Dit kan door een model te bouwen op historische data. Op basis van gemiddelde historische responspercentages, die weergeven hoeveel respons er per periode is binnengekomen, kan al n dagen na de mailing een goede voorspelling voor de totale respons worden gegeven. Ook volgt uit dit onderzoek dat de eerste n weken van de eerste contacten een goed beeld geven van de productie die door de actie getriggered is.

Uit de andere onderzoeken blijkt dat het responsverloop niet gemodelleerd kan worden door zowel tijdreeks modellen als poisson processen. Ook volgt dat *responstijd* een variabele is die niet voorspeld kan worden.

Het bouwen van een tool die de gevonden methode van voorspellen op basis van historische responspercentages implementeert, is een logisch vervolgonderzoek. Het zou ook interessant zijn om te onderzoeken welke factoren *wel* invloed op de variabele *responstijd* hebben.

Inhoudsopgave

1 Inleiding

In de wereld van marketing wil men de resultaten van de campagnes zo snel mogelijk gebruiken voor nieuwe campagnes. Men loopt hierbij tegen het probleem aan dat het vaak lang (meerdere weken) duurt voordat alle reacties van klanten bij de afdeling CI zijn binnengekomen.

Als op basis van de eerste paar dagen respons voorspeld zou kunnen worden hoe de verdere respons zich ontwikkelt, dan zou men veel sneller in de markt kunnen acteren. Dit onderzoek is voor het product hypotheken uitgevoerd. De volgende onderzoeksvraag is geformuleerd:

Kan het responsverloop van het een marketingcampagne voorspeld worden op basis van het begin van de responsperiode?

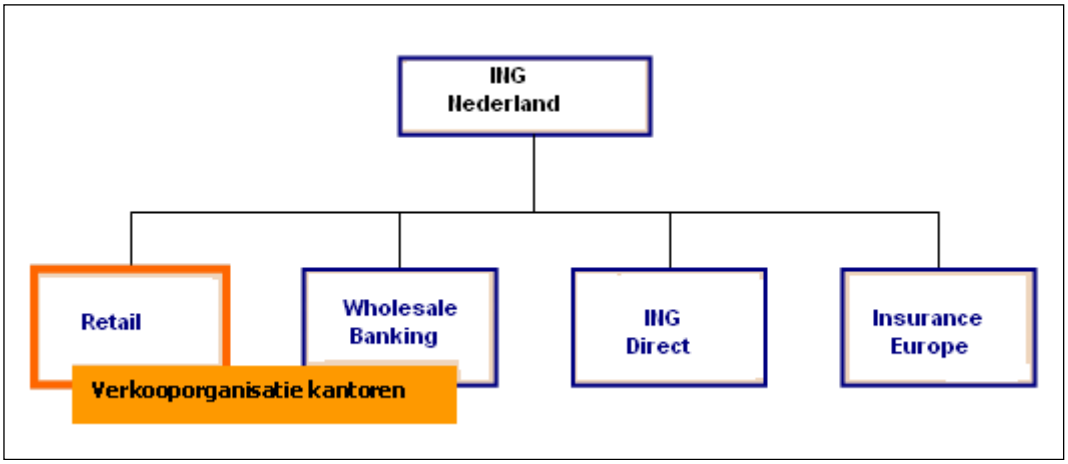
Dit onderzoek is op vier manieren benaderd en bestaat dus ook uit vier onderzoeken. Als eerste is met behulp van *tijdreeks analyse* geprobeerd het responsverloop van de eerste contacten te voorspellen. Een andere benadering is meer op *klant niveau* gericht: *zijn er klanten met bepaalde eigenschappen die altijd vroeg of laat responderen*? Daarna is onderzocht of bepaalde begin dagen goede voorspellers zijn voor de totale respons op basis van gemiddelde historische respons percentages. Als laatste is het responsverloop door een Poisson proces gemodelleerd.

Als eerste wordt het bedrijf beschreven waar de stage heeft plaatsgevonden. Daarna wordt in hoofdstuk 3 een introductie gegeven over het opzetten en evalueren van acties bij de Postbank. In hoofdstuk 4 wordt het onderzoeksprobleem besproken en in hoofdstuk 5 de aanpak van het onderzoek. In hoofdstuk 6 wordt onderzocht in welke periodes de productie zit die door de actie getriggered is. De verschillende benaderingen van het probleem worden in de hoofdstukken 7,8 9 en 10 uitgewerkt. De resultaten van elk onderzoek zijn in hoofdstuk 11 terug te vinden. Tenslotte zijn de voornaamste conclusies en aanbevelingen in hoofdstuk 12 te lezen.

2 ING Retail

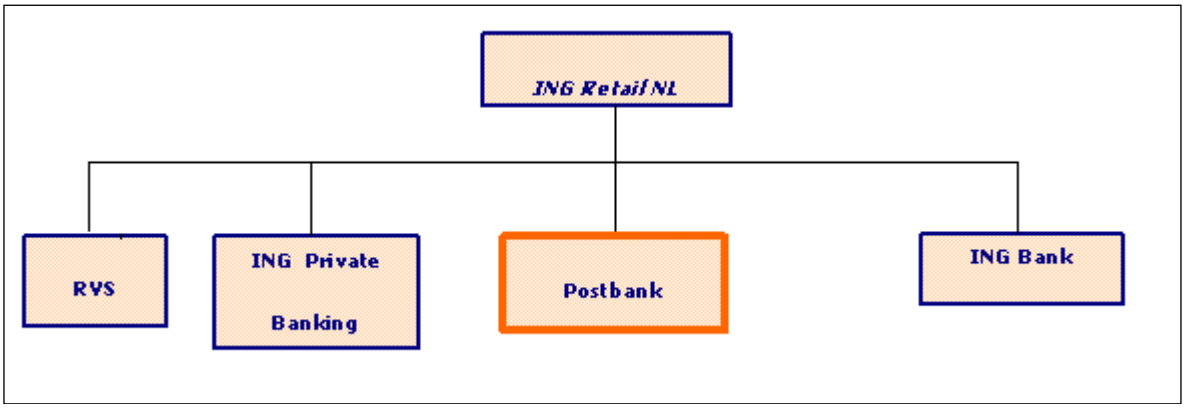
ING Retail/Customer Intelligence/Research & Modelling
Het onderzoek is uitgevoerd op de afdeling Research & Modelling. Dit hoofdstuk beschrijft deze afdeling binnen de ING.

ING Nederland
ING Nederland bestaat uit vier divisies. Het doel van iedere divisie is om de klant service, gemak, zekerheid en overzicht te bieden in zijn financiële situatie. De organogram van ING Nederland en haar divisies zijn in figuur 1 weergegeven.



Figuur 1: *Organogram ING Nederland*

ING Retail NL
De divisie Retail NL is opgericht in 2003 en bestaat uit de labels RVS, Postbank, ING Bank en ING Private Banking.



Figuur 2: *Organogram ING Retail*

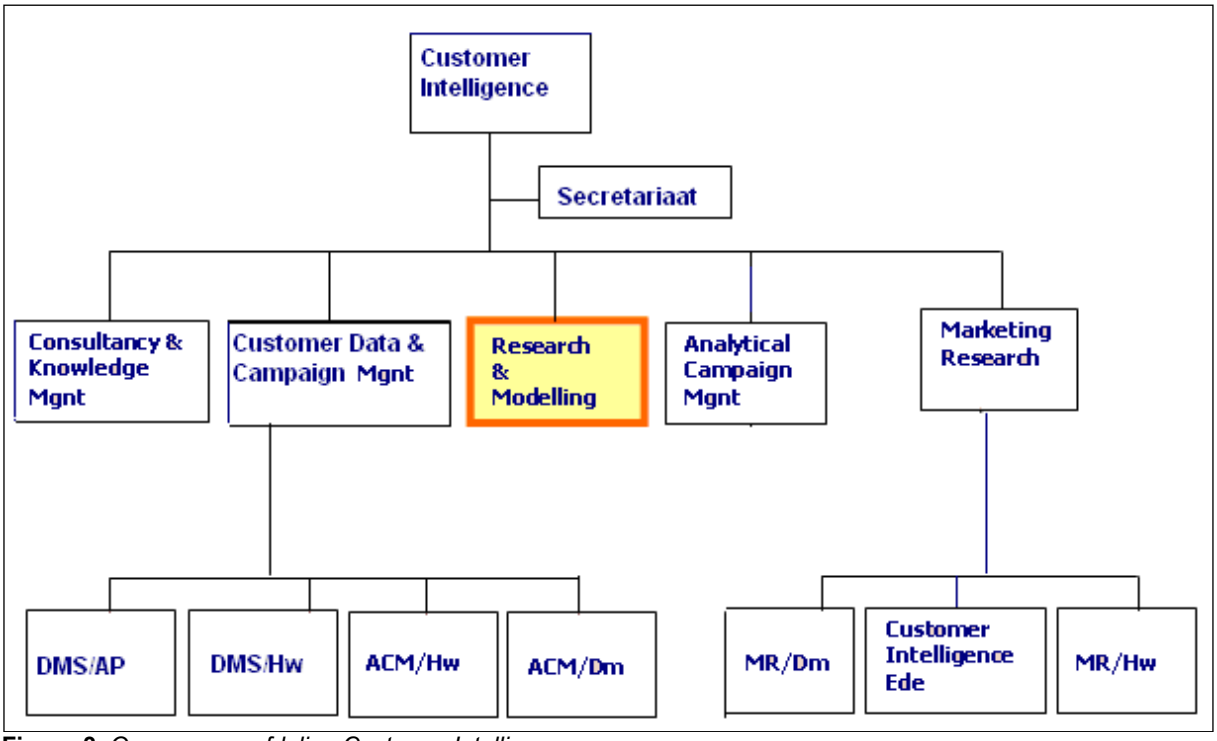
ING

Afdeling Customer Intelligence

Customer Intelligence (CI) is een van de onderdelen binnen ING Retail. CI is verantwoordelijk voor het voeren van de klantregie voor alle klanten van Retail, dat wil zeggen voor de labels Postbank, ING Bank en RVS.
Customer Intelligence ondersteunt op vijf manieren de marktbewerking door de commerciële labels van ING Retail Nederland:

- door het beheren en 'verrijken' van informatie over klanten in de centrale database
- door het ontwikkelen en beschikbaar stellen van klant- en marktkennis uit interne en externe bronnen
- met ondersteuning bij het uitvoeren van campagnes en marketingprojecten
- door het ontwikkelen van een visie en een beleid ten aanzien van klantregie
- door zich constructief kritisch en kostenbewust op te stellen als partner in business, met een maximaal commercieel resultaat als doel.

Customer Intelligence bestaat uit vier onderdelen. Twee daarvan onderhouden directe contacten met opdrachtgevers: Marketing Research en Analytical Campaign Management. De drie andere bieden intern hoogwaardige specialistische ondersteuning. Dit zijn Customer Data & Campaign Management, Research & Modelling en Consultancy & Knowledge Management.



Figuur 3: *Organogram afdeling Customer Intelligence*

Afdeling Research & Modelling
Research & Modelling staat voor hoogwaardige data-analyse, optimale modellen en state-of-the-art advies over methoden en technieken voor onderzoek, modelling, evaluatie en experiment. Het werk van R&M is van directe invloed op:

- selecties voor marketingacties (m.n. door marketingmodellen)
- acceptatie- en beheersprocessen van kredietproducten (m.n. door risicomodellen en/of scorecards)
- de wijze waarop ING Retail evalueert en experimenteert (m.n. door meetplannen, testopzetten en bijbehorende evaluaties).

3 Opzetten en evalueren van marketingacties

Hoe worden acties bij de Postbank opgezet en geëvalueerd? De informatie uit dit hoofdstuk komt uit het boekje *Slim Testen en Vet rendement (Tip W.)* van de Postbank behalve de paragrafen *Bruto en Netto Respons*, *Evaluatietool* en *Conversie percentages*.

Het campagne proces bestaat in hoofdlijnen uit de volgende stappen:

- *Intake*
- *Opzetten benadering*
- *Selectie*
- *Start benadering*
- *Verwerk respons*
- *Evalueer*

Elk marketing campagne begint met een **intake** tussen de afdeling Customer Intelligence, de budgethouder en de uitvoerende partijen over wat er moet gebeuren. Daarna wordt de **benadering uitgewerkt**, waarbij CI een voorstel doet voor doelgroepselectie en een testopzet. Nadat CI klanten **geselecteerd** heeft, kan de **benadering van start.** Een benadering bestaat uit inbound- of uit outboundcontacten[1]. Nadat de backoffice de **respons verwerkt** heeft en het in de database zit, kan CI beginnen met het **evalueren** van de benadering. Hieruit volgt een oordeel over de herhaalbaarheid van de actie. De evaluatie is input voor een **bespreking** van een nieuwe campagne.

In de volgende paragrafen wordt meer uitvoerig op de stappen *testopzet* en *evalueren* ingegaan.

3.1 Testopzet

Een testopzet is een opzet waarbij vooraf gedefinieerde variabelen tegen elkaar worden afgezet om informatie te verzamelen voor een eventuele vervolgactie. Hieronder worden enkele van deze besproken.

[1] Outbound is bv een mailing sturen of een klant bellen
Inbound : Suggesties aan telefoon, Persoonlijke banners

3.1.1 Targets

Voordat een actie wordt uitgevoerd, wordt het doel vastgesteld. Welke klantbehoefte wordt vervuld en welke winstbijdrage levert de actie? Een actie moet een bijdrage leveren aan de algemene MTP-targets[2].

Er zijn een paar belangrijke verschillen tussen MTP-targets en actietargets. Één Belangrijk verschil is dat:

- MTP-targets zijn inclusief de autonome productie, terwijl actietargets alleen over de productie gaan die door een actie wordt veroorzaakt.

Voor het verschil tussen autonoom en actie zie het volgende hoofdstuk over referentiegroepen.

Bij sprake van een *target* in dit onderzoek, wordt de actietarget bedoeld.

3.1.2 Referentiegroepen

Het uitgangspunt bij referentiegroepen is om te leren van de diverse marketingacties die zijn uitgezet. Door een groep op te nemen die de marktbeweging niet krijgt, kan bepaald worden wat de autonome groei is en daarmee het netto resultaat van de actie. Hieronder worden de definities van *bruto* en *netto respons* uitgebreid besproken.

Bruto en Netto Respons

Bij de Postbank wordt gesproken van *netto* en *bruto* respons. *Bruto* respons houdt in dat de gemeten respons bij de actiegroep nog niet vergeleken is met de referentiegroep. Er kan dus nog weinig over het succes van de actie geconcludeerd worden omdat dit aantal nog niet vergeleken is met de referentiegroep.

De *netto* respons wordt verkregen door het responspercentage van de actiegroep af te trekken van het responspercentage van de referentiegroep. Dit procentuele verschil is de *netto procentuele respons*. Door deze te vermenigvuldigen met de grootte van de actiegroep kan de *netto respons* worden bepaald.

Voorbeeld
Stel de actiegroep 5.000 klanten bevat en de referentiegroep 2.500 klanten.
De gemeten respons bij de actiegroep is 400 en bij de referentiegroep is deze 100.
Met deze responsaantallen kan voor beide groepen het responspercentage bepaald worden:

- Actiegroep: 400/5.000= 8% respons
- Referentiegroep: 100/2.500 = 4% respons

Deze 400 respondenten zijn een *bruto* aantal respondenten omdat dit aantal nog niet vergeleken is met de referentiegroep. De 8% is dus een *bruto responspercentage*.
Om de netto respons te krijgen, en dus wat de actie meer heeft opgeleverd dan bij de referentiegroep, wordt het verschil van de responspercentages genomen:
8% respons – 4% respons= 4% netto respons.
De *netto respons* is dan 4% *5.000=200. De actie heeft dus 200 extra respons opgeleverd.

[2] MTP= Middellange termijn planning. In de jaarlijks op te stellen Middellange Termijn Planning (MTP) geeft de divisiedirectie haar visie op de markt en geeft daarmee aan wat we in de komende drie jaar (2006 - 2009) als divisie willen bereiken, hoe we dat gaan bereiken en welke koers we daarbij kiezen

ING

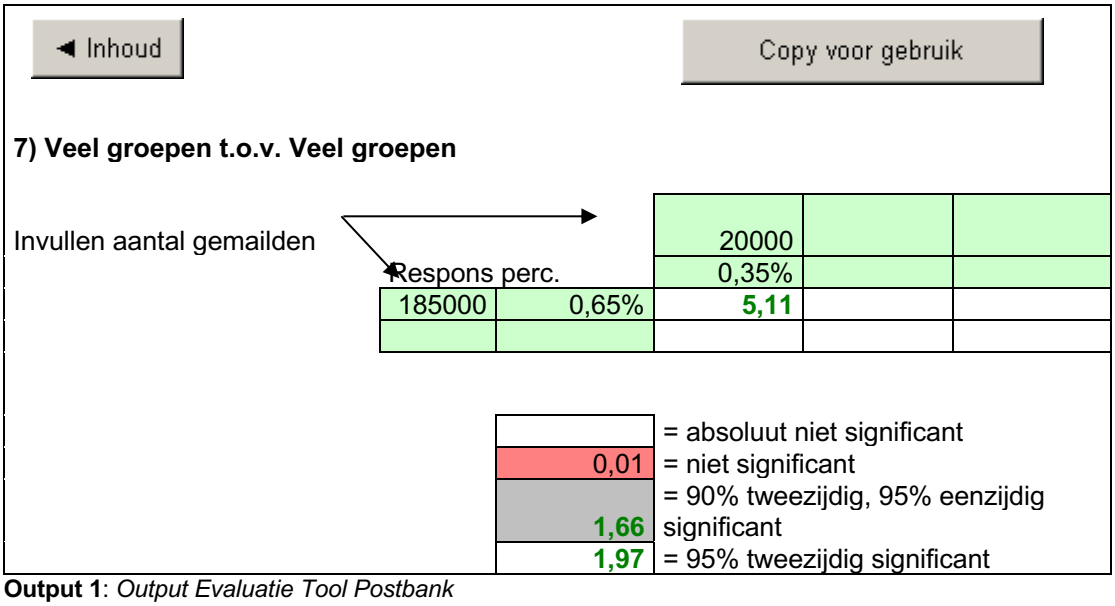
Evaluatietool

De Postbank maakt gebruik van de *Evaluatie Tool* om de verschillen tussen de actiegroep en referentiegroep te evalueren. Dit is een Excel rekenblad waarin handmatig de grootte van de actiegroep en de referentiegroep wordt ingevoerd. Voor elke groep wordt dan tevens het gemeten responspercentage ingevoerd.
De EvaluatieTool berekend dan of de verschillen in responspercentages van de actiegroep en referentiegroep significant zijn of niet.

Het bepalen of deze responspercentages significant van elkaar verschillen wordt gedaan door de *verschiltoets voor fracties*[3] toe te passen.

Voorbeeld
Stel er is een actie geweest waarbij een mailing naar 185.000 klanten is verstuurd. De gebruikte referentiegroep is 20.000 groot. De gemeten respons bij de actiegroep was 1200 eerste contacten en bij de referentiegroep 70. Het responspercentage bij de actiegroep is dan $0{,}65\% = (\frac{1200}{185000})$ en bij de referentiegroep $0{,}35\% = (\frac{70}{20000})$.

Als er een significant verschil is tussen deze responspercentages betekent dit dat de actie dus significant meer respons heeft behaald dan de referentiegroep. In output 1 is de output van de Evaluatie Tool te zien. Hieruit volgt dat er de responskans van de groepen significant verschillend is. Deze actie heeft dus significant meer respons behaald dan de referentiegroep. Dit is wat van een actie natuurlijk ook verwacht wordt.



Output 1: *Output Evaluatie Tool Postbank*

[3] De theorie van deze toets wordt in hoofdstuk 6.*1.2* besproken

3.1.3 Selectietesten

Voor een goede direct marketing campagne is het maken van de juiste selectie van klanten erg belangrijk. Uit ervaring blijkt dat het zelf bedenken van een selectie (common sense) meestal niet tot optimale resultaten leidt. Wat wel werkt is ervaring. De beste manier om ervaring met selecties op te bouwen is doelbewust testen. De beste test is een randomgroep (een willekeurige selectie binnen alle klanten die het product juridisch mogen afnemen) gevolgd door modelbouw. Bij modelbouw gebruikt de Postbank statistiek om te bepalen wat het profiel is van klanten die responderen. Vervolgens worden met het model alleen die klanten aangeschreven die een hoge responskans hebben. In dit hoofdstuk worden verschillende voorbeelden van selectietesten besproken.

Modellen

Een model is een verzameling selectiecriteria waarmee je klanten kunt indelen naar responsgeneigdheid voor een benadering. Met een model kunnen de beste klanten voor een benadering uit de database gehaald worden.

Stappen modelbouw

1. doe een random-selectie;
2. bouw model=kijk welke kenmerken respondenten anders hebben dan non respondenten. Die informatie noemen we het model;
3. pas het model toe op alle klanten. Nu gaan de klanten die het aanbod nog niet gehad hebben een score krijgen. Klanten met ‘goede’ kenmerken krijgen een positieve score;
4. selecteer de goede klanten voor actie.

3.2 Uitvoer evaluatie

Na te hebben bepaald wat je wilt meten bij de actie-opzet, is het vervolgens zaak de opzet te evalueren. Het onderwerp evalueren van acties en testopzetten wordt nu verder uitgewerkt.

3.2.1 Van een actie tot respons

Bij een organisatie met een multichannel communicatie strategie probeert men de klant niet alleen via meerdere kanalen te benaderen maar heeft de klant ook de mogelijkheid om via verschillende kanalen te responderen. Deze werkwijze heeft gevolgen voor de evaluatie van acties en van kanalen. Voor het goed begrijpen van multichannel is het van belang een onderscheid te maken tussen benaderingskanaal en responskanaal.

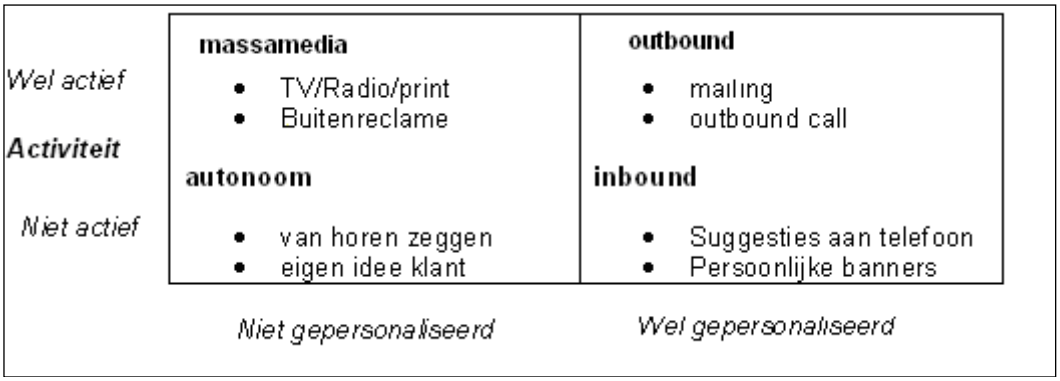
Een benaderingskanaal geeft aan hoe de klant op het idee komt om een product bij ons af te nemen. Het responskanaal geeft aan op welke manier de klant het product bij ons afneemt.

Voor het benaderingskanaal en responskanaal gelden verschillende analysetechnieken. Zo kan voor het benaderingskanaal bepaald worden welke respons dit kanaal veroorzaakt (door vergelijking met de referentiegroep). Voor het responskanaal kan bekeken worden hoeveel klanten er in de verschillende

responstappen afvallen. De opdeling van respons in responstappen wordt ‘waterval’ genoemd.

Benaderingskanaal

In figuur 4 staat een aardige manier om tegen een benaderingkanaal aan te kijken. De Postbank kan de klant actief benaderen, via massa of direct media. De Postbank kan ook wachten tot de klant bij hun komt. Als de Postbank van te voren bedacht heeft wat ze doen als de klant bij hun komt, spreken we van gepersonaliseerde inbound. Komt de klant helemaal uit zichzelf, dan spreken we van autonome groei.



Figuur 4: *Benaderingskanalen*

Responskanalen en waterval

Responskanalen kunnen op twee manieren bekeken worden:

1. Respons als gevolg van een benadering
Dit is de traditionele direct marketing meetmethode. Hierbij wordt de respons gerelateerd aan de benaderende klanten en vergeleken met een referentiegroep. Doe je dit per responskanaal, dan krijg je voor een benadering een multichannel-respons evaluatie.

2. Doelmatigheid van responsverwerking (funnel/waterval)
Als de analyse vanuit het responskanaal uitgevoerd wordt, dan kan gekeken worden naar de verschillende responsstappen. Door per stap te kijken hoeveel klanten er overblijven (conversie), bepaal je de effectiviteit van het responskanaal. Deze manier van kijken wordt de watervalmethode genoemd. Voordeel van de watervalmethode is dat je veel inzicht krijgt over de verwerking van de respons.

Conversie percentages

Een conversie percentage is een percentage van bijvoorbeeld eerste contacten die effectief leads[4] worden.Bij de Postbank maken ze gebruik van vaste conversie percentages die door onderzoeken binnen Postbank zijn vastgesteld.

3.2.2 Snelheid van evalueren

Het moge duidelijk zijn dat alle betrokken partijen zo snel mogelijk willen weten wat de vruchten van de arbeid zijn. Hoe eerder de Postbank het weet, hoe eerder de Postbank kan bijsturen of reageren. De snelheid van evalueren heeft veel aspecten die in de volgende paragrafen worden uitgewerkt.

Monitoring, evaluatie en snelheid

Het meten van de resultaten van een campagne heeft twee doelen:

- *Bijsturen*
  Informatie voor bijsturen kan gaan over het voorraadbeheer van premiums, capaciteitsplanning van agents/adviseurs of het al dan niet halen van targets. Doel is bijsturen, waarbij tijdigheid van de op te leveren informatie zeer belangrijk is. Het verzamelen van dit soort informatie wordt *monitoren* genoemd.

- *Beslissen herhalen ja/nee*
  Informatie voor de herhalingsbeslissing is bij een periodieke actie vaak pas later nodig. Dat is gunstig want dan kan informatie over de actie verzameld en geïnterpreteerd worden (evalueren).

Onder monitoring valt de meting van leads, afspraken, bezoeken en respons. Deze metingen vinden continu plaats. De hoeveelheid informatie die we die van de actie bekend is, tijgt naarmate de tijd verstrijkt. Met de productie in de database en de referentiegroep gemeten, kan de evaluatie afgerond worden. De kennis over de actie stijgt door de evaluatie opeens zeer.

Soms moet het sneller. Het besluit om de actie te herhalen kan niet altijd wachten tot de evaluatie gedaan is. De informatie uit de monitoring moet dan gebruikt worden om een inschatting te maken hoe de evaluatie zal uitpakken. Twee situaties waarbij dit veilig kan:

- *Tegenvallend resultaat*
  Bij acties die ver onder de verwachting presteren, wordt al tijdens het monitoringproces duidelijk dat de targets niet gehaald gaan worden. Het bekijken van de referentiegroep zal nooit meer respons opleveren. Bij een slecht lopende actie kan dus al voor de evaluatie besloten worden dat herhaling niet nuttig is.
- *Positief resultaat*
  Als de Postbank al vaker een bepaal aanbod heeft gedaan, ontstaat er een verwachting over de productie in de referentiegroep. De verhouding respons-netto productie uit een evaluatie van een vorige keer kan worden gebruikt om in te schatten wat de netto productie deze keer zal worden. Is de ingeschatte netto productie hoog genoeg, dan is een herhaling verantwoord.

[4] Een *lead* zit tussen een *eerste contact* en *eerste gesprek.* De klant zoekt contact via Internet, telefoonbank of schriftelijk. De telefoonbank agent belt vervolgens de klant voor een afspraak en boekt in. Dit is een *lead*.

Verdeling van de respons in tijd

De tijd tussen de actiedatum en datum waarop begonnen kan worden met de evaluatie is met name afhankelijk van de referentiegroep. De klanten moeten namelijk de tijd krijgen om te responderen.

Het gebruiken van de juiste meetperiode is belangrijk voor het trekken van de juiste conclusie. Wat de juiste meetperiode is, hangt af van de product-klantgroep combinatie. Voor de vergelijkbaarheid van evaluaties is in ieder geval belangrijk dat elke keer dezelfde tijdsduur wordt gebruikt.

4 Probleem beschrijving

De probleem beschrijving bestaat uit het formuleren en beschrijven van een aantal begrippen. Deze worden in de volgende paragrafen beschreven.

4.1 Doel Onderzoek

Doel algemeen
In de wereld van marketing wil men de resultaten van de campagnes zo snel mogelijk gebruiken voor nieuwe campagnes. Men loopt hierbij tegen het probleem aan dat het vaak lang (meerdere weken) duurt voordat alle reacties van klanten bij de afdeling CI zijn binnengekomen.
Als op basis van de eerste paar dagen respons voorspeld zou kunnen worden hoe de verdere respons zich ontwikkelt, dan zou men veel sneller in de markt kunnen acteren.

Omdat de responsperiode voor *hypotheken* standaard n dagen (n maanden) is, kan de actie pas na n maanden geëvalueerd worden. Voor dit product is er dus veel voordeel bij als de actie vervroegd geëvalueerd zou kunnen worden. Hieronder wordt het doel voor het product hypotheken besproken.

Doel product hypotheken
Wat wil de Postbank met dit onderzoek bereiken voor het product *hypotheken?*
Als vroeg in de actie al geëvalueerd kan worden hoeveel hypotheken de actie gaat opbrengen, en dus hoe effectief een actie is, dan kan redelijk vroeg in de actie worden ingezien of een actie genoeg hypotheken zal gaan opleveren, en dus of deze winstgevend zal zijn of niet. Als blijkt dat de target op hypotheken productie niet gehaald kan worden, dan kan worden besloten de actie aan te passen en zo uitgaven te beperken of met extra uitgaven alsnog de target te halen.

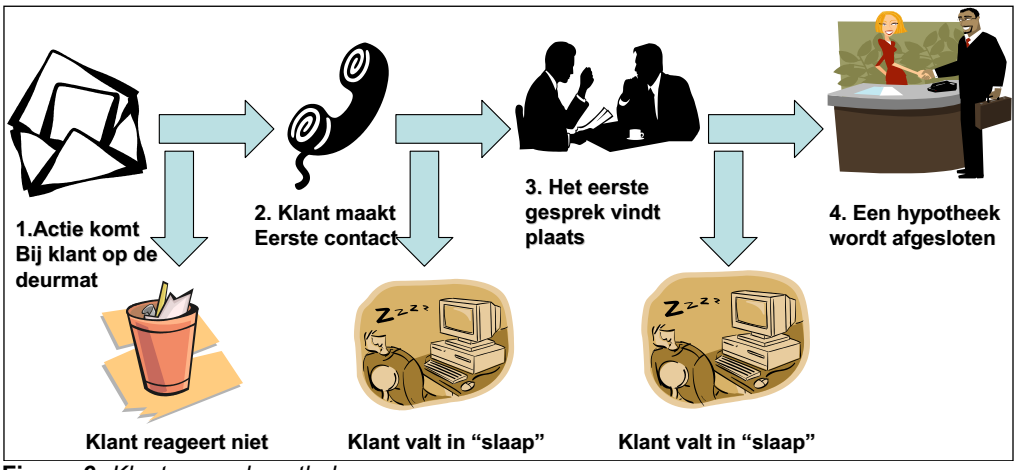
4.2 Verkoopproces Hypotheken

Hoe verloopt het verkoopproces bij de Postbank voor het product hypotheken?
Het verkoopproces wordt vanuit de klant bekeken worden en wordt hieronder beschreven.

Klantproces
Het klantproces bestaat uit enkele stappen. Het begint op het moment dat een klant een brief over een actie ontvangt. Een klant kan hierop besluiten te reageren of niet. Als hij besluit te reageren/responderen wordt dat als een **eerste contact** geregistreerd.
Een klant kan op verschillende manieren responderen. Hij kan dit doen door:

- een coupon in te vullen,
- via internet een formulier in te vullen,
- via de telefoon met de Postbank Advies Lijn,
- via het postkantoor bij de Postbank Advies Balie.

De telefoonbank agent belt vervolgens de klant voor een afspraak en boekt in. Dit is de ***lead***. Daarna gaat de adviseur het gesprek aan met de klant, **het eerste gesprek**. Als alles goed verloopt, maakt de adviseur een hypotheekofferte voor de klant. Als de klant akkoord gaat met het voorstel, wordt een hypotheek afgesloten. Dit is de **openingsdatum**.
Bij elke stap kan een klant “in slaap” vallen. Dat houdt in dat een klant vergeet of geen interesse meer heeft om verder te gaan. Dit proces is beschreven in figuur 6.



Figuur 6: *Klantproces hypotheken*

4.3 Beschikbare data

Om een goed en betrouwbaar onderzoek te kunnen doen, moeten verschillende acties gebruikt worden om zo een beter beeld te krijgen van het responsverloop. Er is grote hoeveelheid aan hypotheken data beschikbaar. Wegens het vele werk aan datapreparatie dat per actie gedaan moet worden en de beschikbaarheid van deskundigen voor het zorgen van deze bestanden, heb ik met drie acties gewerkt; actie A6144, actie A5727_B en actie A5727_C. Dit zijn zogenaamde *Oversluit acties*. Dit zijn campagnes die klanten werven om hun hypotheek over te laten sluiten bij Postbank. Het zijn acties die in 2005 hebben plaatsgevonden.

4.4 Oplevering Onderzoek

Dit onderzoek moet leiden tot meer kennis over waar de hypotheek productie van de actie zit. Ook moet dit onderzoeken leiden tot een methode die gebruikt kan worden om de productie van een actie te voorspellen. Deze methode kan slechts opgesteld worden als blijkt dat de productie van een actie voorspeld kan worden.

5 Probleem aanpak

Dit onderzoek begint met een data mining proces om zo tot een beter inzicht van het probleem te komen. Daarna volgen enkele *Modelleer Technieken* waarmee het probleem aangepakt gaat worden.

5.1 Data mining proces

Hier wordt een beeld geschetst van hoe het probleem aangepakt gaat worden en welke datagegevens hiervoor nodig zijn. Daarna wordt de data verzameld. Dan volgt een fase waarin de data goed geanalyseerd moet worden om zo mogelijke fouten in de data op te sporen. Voor gevonden fouten worden dan mogelijke verklaringen gevonden.

5.1.1 Probleem begrip

Het doel van dit onderzoek is om de productie te kunnen voorspellen die een actie opbrengt. Bij de Postbank is het percentage eerste contacten dat leidt tot het afsluiten van hypotheken bekend. Dit percentage wordt een *conversie* percentage genoemd. Het idee is dus om op basis van de eerste paar dagen van de eerste contacten, het totaal aantal eerste contacten te voorspellen. Als dit totaal aantal eerste contacten voorspeld kan worden dan kan met behulp van het conversie percentage de productie bepaald worden die uit deze contacten volgt. We willen dus het totaal aantal eerste contacten gaan voorspellen. De benodigde datagegevens voor het product hypotheken zijn dus de *eerste contact* gegevens en de *openingsdatums*.

5.1.2 Dataverzameling

Om een onderzoek uit te voeren moet data verzameld worden. Deskundigen bij de Postbank hebben historische data van enkele acties verzameld en deze in SAS bestanden klaargemaakt. Uiteindelijk zijn bestanden van drie acties klaar gemaakt: actie A6144, actie A5727_B en actie A5727_C.

5.1.3 Verklarende data analyse

Voor aan onderzoek beginnen is het van belang de data goed te analyseren en zo mogelijke fouten in de data op te sporen en mogelijke verklaringen hiervoor te vinden. Als eerste wordt een structuur van elke actie gemaakt die toont hoeveel eerste contacten, gesprekken en openingsdatums elke actie bevat. Daarna worden de fouten in de data besproken en mogelijke verklaringen hiervoor gevonden.

5.1.3.1 Structuur data

5.1.4 Model ontwikkeling & validatie

Voor het maken van een model wordt de dataset gesplitst in twee datasets, namelijk de training set en de test set. De verhouding is meestal 75% voor de training set en 25% voor de test set. Na het model gebouwd te hebben is het van belang het model te testen. Dit wordt op de test set gedaan. Deze techniek wordt voornamelijk gebruikt bij classificatie technieken maar kan ook bij andere technieken gebruikt worden. Bij de modelleer techniek *Gemiddeldes* wordt een model gebouwd op twee acties en wordt het verkregen model getest op een derde nieuwe actie. Op deze manier kan een inzicht worden verkregen van de kwaliteit van het model.

5.2 Modelleer technieken

Dit onderzoek is op vier manieren benaderd en bestaat dus uit vier onderzoeksdelen. Hieronder wordt elke techniek die per deel gebruikt is besproken.

5.2.1 Tijdreeks analyse

Met behulp van tijdreeks analyse wordt geprobeerd het responsverloop van de eerste contacten te modelleren. Er wordt gezocht naar een proces die voor alle beschikbare acties geschikt is en het responsverloop modelleert. Op basis van het gevonden model kan dan voor een nieuwe actie het responsverloop voorspeld worden.

5.2.2 Lineaire regressie

Een andere aanpak is onderzoeken of klanten met bepaalde eigenschappen altijd vroeg of laat responderen. Een statistische methode om te onderzoeken hoe de *responstijd* variabele afhangt van een of meerdere klant variabele is *regressie*. Bij lineaire regressie wordt de relatie tussen de betrokken variabelen weergegeven door een lineaire vergelijking. Met behulp van deze techniek wordt geprobeerd een vergelijking te vinden die de variabele *responstijd* het beste benadert. Op basis van de opgenomen variabelen in de model vergelijking kan dus voor een bepaalde klant worden bepaald wat zijn/haar responstijd is.

5.2.3 Poisson proces

Door het proces van het binnenkomen van de respons (eerste contacten) als een poisson proces te modelleren wordt geprobeerd de totale respons te voorspellen. Elke actie heeft een eigen parameter λ die voor elke actie geschat wordt op basis van de data van die actie alleen. Dit kun je na 1 dag doen, na 2 dagen, na 3 dagen etc. Naar verwachting wordt de voorspelling steeds beter, als het model redelijk is. Na elke dag kan de parameter van het poisson proces beter geschat en kan dus het totaal aantal eerste contacten steeds beter voorspeld worden.

5.2.4 Gemiddeldes

Op basis van historische data worden gemiddelde responspercentages genomen van de waargenomen responsaantallen die per periode binnengekomen zijn. Met periode wordt *eerste dag, eerste twee dagen* etc. bedoeld. Voor een nieuwe actie kan dan op basis van deze percentages een voorspelling worden gemaakt van de totale respons. Op basis van het testen van dit model op een andere actie kan bepaald worden vanaf welke periode een goede voorspelling voor de totale respons gedaan kan worden. Voor een nieuwe actie kan dan na die bepaalde periode een schatting worden gemaakt voor de totale respons. Dit door de modelpercentages te gebruiken die gemaakt zijn op de historische data.

6 Bepalen productie actie

Voor aan het onderzoek te kunnen beginnen is het van belang om te onderzoeken in welke periode de productie van de actie zit. Dit is namelijk wat men wil kunnen voorspellen. In dit hoofdstuk wordt dit vooronderzoek, naar het bepalen van de productie die door de actie is getriggered, besproken.

Het bepalen in welke periode de hypotheken productie zit van een actie kan om twee manieren onderzocht worden. Als eerste manier kan bepaald worden in welke periode de eerste contacten zitten die significant hoger zijn dan bij de referentiegroep en dan de productie bepalen die uit deze contacten komen. Deze eerste contacten zijn dan een effect van de actie en dus is de hypotheek productie die hieruit volgt de productie die door de actie komt. Dit wordt de *Waterval* methode genoemd.

Een tweede manier is het bepalen van de periode waar de hypotheken productie significant hoger is dan bij de referentiegroep. Het relatieve verschil tussen de productie van de actiegroep en de productie van de referentiegroep zou dus de productie zijn die door de actie komt. Dit wordt de *Directe* methode genoemd. Deze twee methodes zullen beide besproken en behandeld worden.

In dit hoofdstuk wordt eerst de theorie over *toetsingsprocedure*, *de binomiale verdeling* en *bootstrappen* besproken. Daarna wordt voor het product *hypotheken* op de twee besproken manieren de periode bepaald waar de productie van de actie zit.

6.1 Theorie

6.1.1 Toetsingprocedure

Een onderzoek moet leiden tot een samenvattende uitspraak of conclusie. Daarin wordt een antwoord gegeven op de gestelde onderzoeksvraag. Die uitspraak wordt meestal gedaan op grond van steekproefonderzoek, maar betreft de populatie, waaruit de steekproef is getrokken.

Wetenschappelijke conclusies hebben veelal de vorm van een statistische toets. In een statistische toets wordt eerst een zo concreet mogelijke *(nul)hypothese* geformuleerd over de populatie en vervolgens wordt de juistheid van deze hypothese getoetst. Het resultaat van de toets is de uitspraak, dat de hypothese op grond van de resultaten van de steekproef kan worden verworpen, of juist niet kan worden verworpen. Kenmerkend voor een statistische toets is dat wordt aangegeven hoe groot de kans is dat de hypothese ten onrechte wordt verworpen.

Bij een toetsingsprobleem worden de volgende stappen genomen:

1. Formuleer een nulhypothese (H_0) en een alternatieve hypothese (H_1).
2. Kies een waarde voor α, dit is het significantieniveau oftewel fout van de eerste soort. Bij geldigheid van de nulhypothese mag er voor de steekproefgrootheid een kans zijn op een uitkomst in het kritieke gebied *Z*, die hoogstens α bedraagt.
3. Bepaal met welke toetsingsgrootheid er gewerkt moet worden.
4. Bereken het kritieke gebied *Z*.
5. Bepaal de uitkomst van de toetsinggrootheid en bekijk met behulp van het kritieke gebied *Z* of de nulhypothese al dan niet verworpen moet worden. Als beslisregel geldt:
    - Als de gevonden waarde in *Z* ligt dan wordt H_0 verworpen,
    - Als de gevonden waarde niet in *Z* ligt dan wordt H_0 aangenomen.
6. Geef een formulering van de conclusie.

Hieronder volgen de definities *van hypothese*, *toetsingsgrootheid*, *kritiek gebied Z*, *z-waarde* en het *significantie niveau*.
Deze informatie komt uit *Statistiek om mee te werken* [Buijs (1999)].

Hypothese
De hypothese die we formuleren en toetsen in een statistische toets wordt de nulhypothese H_0 genoemd. De nulhypothese moet bij het begin van het onderzoek worden geformuleerd. Deze hypothese heeft de vorm van een *exacte*, *kwantitatieve* uitspraak over een parameter van de populatie, waaruit de steekproef is getrokken. De alternatieve hypothese (H_1) heeft de vorm van een ontkenning van de nulhypothese. In formule is dat:

H_0: populatieparameter = waarde

H_1 :populatieparameter ≠ waarde

Op deze manier geformuleerd kan de waarde van de populatieparameter zowel groter als kleiner zijn dan de waarde onder de nulhypothese. Omdat de populatiewaarde zowel groter als kleiner dan de *waarde* kan zijn, wordt er *tweezijdig getoetst*. Een meer specifieke alternatieve hypothese is bijvoorbeeld:

H_1: populatieparameter > waarde

Nu wordt er *rechtseenzijdig* getoetst (populatiewaarde kan alleen > waarde zijn). De vorm van de alternatieve hypothese bepaalt dus of er *eenzijdig of tweezijdig getoetst* wordt.

Toetsinggrootheid
De populatie wordt onderzocht door het nemen van een steekproef. Als een steekproef genomen is, dan moeten de verkregen resultaten geanalyseerd worden. Er wordt dan een grootheid opgesteld waarmee de toets wordt uitgevoerd. Dat is de zogenaamde *toetsingsgrootheid*. De bedoeling is dat de informatie uit de steekproef zo goed mogelijk tot uitdrukking komt in een waarde van de toetsinggrootheid.

Veel gebruikte toetsingsgrootheden zijn het steekproefgemiddelde $\bar{x}$ en de steekproeffractie $p = \frac{k}{n}$. Over deze fractie is in paragraaf 6.1.3 meer informatie over te vinden.

Kritiek gebied en voorspellingsinterval

Door een 95 % voorspellingsinterval voor de toetsingsgrootheid op te stellen, kan worden gezegd dat de toetsingsgrootheid met kans 0.95 een waarde in dat interval zal aannemen. Deze collectie uitkomsten worden als normaal 'gekwalificeerd' gegeven het feit dat H_0 juist is. Wat dan overblijft, is een kans 0.05. Deze kans wordt bij een toets aangegeven met α. Dit is dus een de kans om een waarde buiten het voorspellingsinterval te vinden, in de situatie dat de nulhypothese correct is. Als in dat gebied een uitkomst aangetroffen wordt, dan wordt H_0 verworpen.
De verzameling uitkomsten die leiden tot verwerpen van de nulhypothese wordt het kritieke gebied *Z* genoemd. De overige uitkomsten, die dus niet leiden tot verwerping van H_0, duidt men aan als het acceptatiegebied.

In termen van de standaard normale verdeling luidt het kritieke gebied:

$Z=\{x|x<\mu-z\sigma$ of $x>\mu+z\sigma\}$ bij *tweezijdig toetsen*

$Z=\{x|\ x>\mu+z\sigma\}$ bij *rechtseenzijdig toetsen*

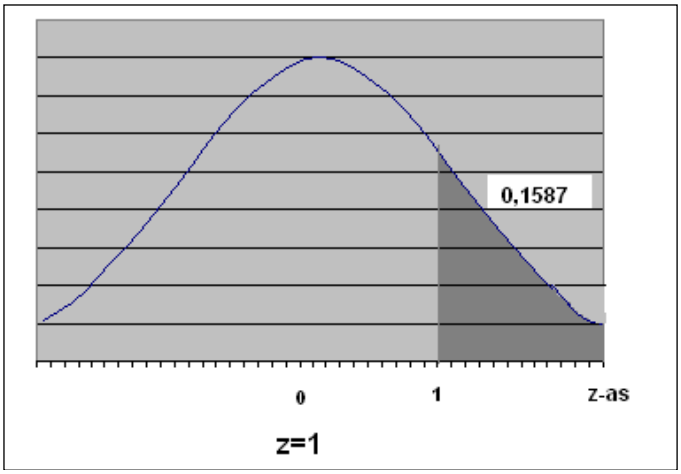
Waarbij *μ* de verwachting van de toetsingsgrootheid, *σ* de standaarddeviatie van de toetsingsgrootheid en *z* de z-waarde is. Het begrip *z-waarde* wordt in de volgende paragraaf besproken. Als de gevonden waarde in *Z* ligt dan wordt H_0 verworpen. Als de gevonden waarde niet in *Z* zit dan wordt H_0 aangenomen.

Z-waarde

De normale verdeling met *μ=0* en *σ=1* noemt men de standaard normale verdeling. Een variabele *z* die een standaard normale verdeling volgt, schrijft men als:

$z\sim N(\mu=0, \sigma=1)$

In de literatuur zijn tabellen bekend van de standaard normale verdeling waar de rechteroverschrijdingskans uit gelezen kan worden. In grafiek 1 is te zien hoe voor *z*=1 (de *grenswaarde*) de kans 0,1587 toebehoort (deze kans is ook uit de tabel van de standaard normale verdeling bij *z-waarde* 1 af te lezen). Dit is de kans om en waarneming te doen die groter is dan 1. De oppervlakte van het staartgedeelte onder de curve ter rechterzijde van de grenswaarde *z=1* bedraagt dus 0,1587.



Grafiek 1: *Kans bij z=1*

Voor een normale verdeling met willekeurig normaal verdeelde μ en σ moet een transformatie worden uitgevoerd. Deze transformatie is een voorbeeld van *standaardisatie* of *normalisatie* van de data. Door standaardisatie zijn de waarden van verschillende variabelen beter te vergelijken. Ook kan door deze transformatie gebruik worden gemaakt van de tabel van de standaard normale verdeling waar vervolgens de kansen opgezocht kunnen worden.
Voor een grenswaarde g in het oorspronkelijke probleem leidt dit tot een grenswaarde in de standaard normale verdeling van:

$$z = \frac{g - \mu}{\sigma}$$

Dit wordt meestal aangeduid als de *z-waarde* die bij een bepaalde grens g hoort. Voor deze z-waarde kunnen vervolgens kansen opgezocht worden in de tabel van de standaard normale verdeling.

Significantieniveau

Om tot een verdeling te komen in een kritiek gebied en een acceptatie gebied is een criterium nodig. Dit wordt gegeven door een kans *α*. Dit wordt *het betrouwbaarheids niveau* of *fout van de eerste soort* genoemd. Deze wordt vooraf gekozen. Bij geldigheid van de nulhypothese mag er voor de steekproefgrootheid een kans zijn op een uitkomst in het kritieke gebeid *Z*, die hoogstens α bedraagt.

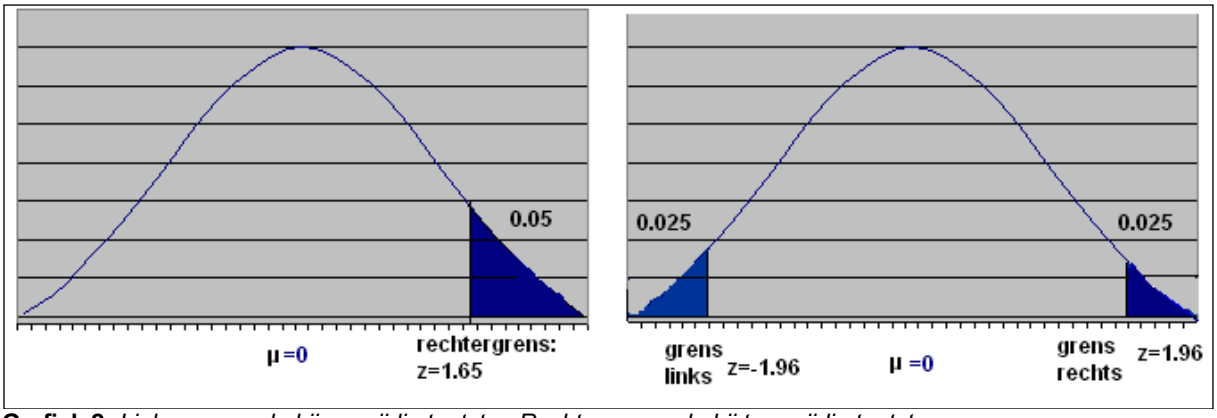
Als eenmaal de waarde voor *α* vaststaat, is het betrekkelijke eenvoudig om voor een kansverdeling aan te geven wat het kritieke gebied *Z* is. Als de steekproef een uitkomst toont die tot *Z* behoort, wordt de nulhypothese verworpen. In formule wordt α weergegeven door:

$$P(x \in Z \mid H_0) < \alpha$$

Dus *α* is de kans om ten onrechte de nulhypothese te verwerpen, want de onderzochte variabele heeft bij geldigheid van H_0 een kans van hoogstens *α* op een uitkomst in *Z*. Een significantieniveau van *α = 0.05* is gebruikelijk en wordt in dit onderzoek ook gebruikt. Het houdt in, dat gemiddeld genomen één op de twintig keer de nulhypothese ten onrechte wordt verworpen.

Keuze z in formule van het kritieke gebied

Bij rechtseenzijdig toetsen levert de waarde $\alpha = 0.05$ een *z*-waarde van *1.65*. Dit volgt uit de tabel van de normale verdeling. Bij tweezijdig toetsten moet aan beide kanten van de verdeling een gebied van met oppervlakte 0.025 gezocht worden (zie grafiek 2). De waarde van 0.025 levert $z=1.96$ in de tabel van de normale verdeling. In de formule voor het opstellen van het kritieke gebied moet dus afhankelijk van of er eenzijdig of tweezijdig getoetst wordt deze waardes voor *z* genomen worden.



Grafiek 2*: Links: z-waarde bij eenzijdig toetsten Rechts: z-waarde bij tweezijdig toetsten*

P- waarde

Bij elke toets kan ook een overschrijdingskans berekend worden. De p-waarde of overschrijdingskans is de kans dat in de verdeling gegeven door de nul-hypothese de waarde van de toetsinggrootheid wordt overschreden. In dit onderzoek wordt de waarde van 5% aangehouden als grens (α); is de p-waarde kleiner dan 5%, dan spreekt men van een **significante** uitkomst en wordt de nul hypothese verworpen.

6.1.2 Binomiale verdeling

De binomiale verdeling is een discrete kansverdeling die een beschrijving geeft van het aantal successen dat kan optreden wanneer men een experiment een gegeven aantal keer (*n*) herhaalt. Bij elke herhaling moet er een vaste kans π zijn op het waarnemen van een succes. Deze herhalingen zijn onafhankelijke gebeurtenissen.

Het aantal toepassingsmogelijkheden van de binomiale verdeling is groot. In de praktijk blijken veel situaties om te vormen tot een binomiale kansverdeling. Het gaat erom dat men te maken heeft met een verschijnsel waarbij de gang van zaken bij een experiment kan worden beschouwd als successen tellen. Het te onderzoeken verschijnsel kan daarbij precies twee toestanden hebben, namelijk succes of pech, 1 of 0, responderen niet responderen. Altijd is het dus een tweedeling.

Binomiale kansformule

Een kansvariabele $\underline{k}$ die een binomiale verdeling volgt, wordt als volgt genoteerd:

$\underline{k}$ ~Bin(n, π)

De binomiale verdeling wordt gekenmerkt door twee parameters, namelijk n, het aantal pogingen, en π , de succeskans per poging.

Algemeen geldt bij een binomiale verdeling: gegeven is een experiment met n pogingen. Per keer bedraagt de kans op succes π en de kans op pech 1- π .
Stel $\underline{k}$=*aantal successen*. We zoeken $P(\underline{k}=k)$. Bij n pogingen zoeken we derhalve naar de kans op k successen en $n-k$ pechgevallen. De succeskans π moet zich dus k maal voordoen en de pechkans $(1-\pi)$ moet zich $n-k$ maal voordoen. Voor elk rijtje van k successen en $n-k$ pechgevallen vinden we als kans $\pi^k(1-\pi)^{n-k}$.

Het aantal gunstige rijtjes van k successen en $n-k$ pechgevallen bedraagt in dit geval $\binom{n}{k}$.
Voor de kans op k successen bij n pogingen vinden we dus:

$$P(\underline{k}=k)=\binom{n}{k}\pi^k(1-\pi)^{n-k} \qquad (1)$$

Dit is de *algemene formule* voor de binomiale verdeling.

Verwachting en Variantie

Voor de binomiale verdeling is de verwachting:

$$E(\underline{k})=\sum kP(\underline{k}=k)=n\pi$$

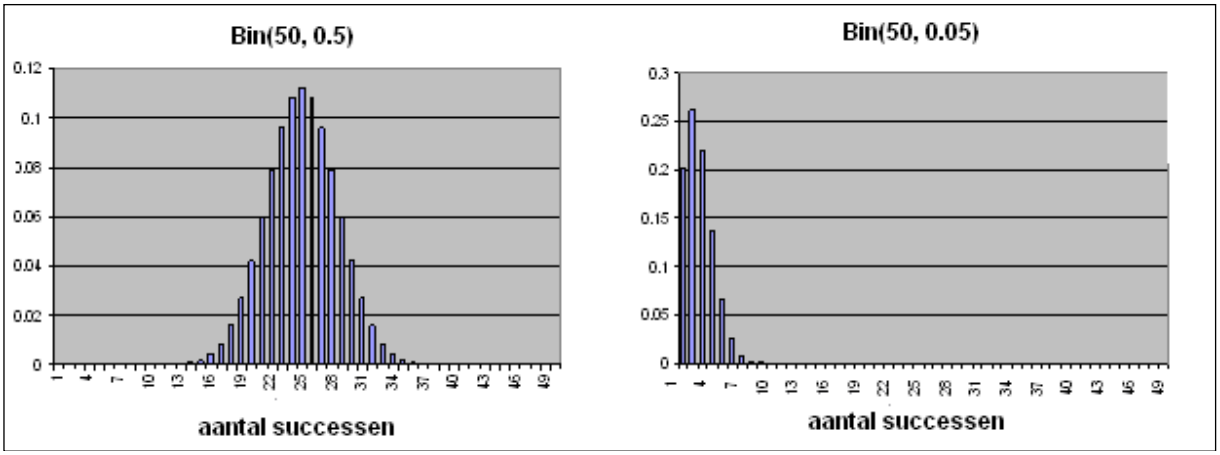
en de variantie:

$$Var(\underline{k})=n\pi(1-\pi)$$

De Normale of Poisson benadering

Indien de steekproefomvang n heel groot is, kan of de normale verdeling of de poisson verdeling gebruikt worden om de kansen uit de binomiale verdeling te berekenen. Deze verdelingen hebben de voorkeur omdat als de steekproefomvang n heel groot is, er rekentechnische moeilijkheden dreigen. Niet alleen wordt de uitdrukking $\binom{n}{k}$ lastig te berekenen maar ook de het getal $(1-\pi)$ zal tot een hoge macht moeten worden gebracht.
Voor beide methodes geldt dat zij slechts toegepast mogen worden indien de steekproefomvang n voldoende groot is. Als criterium wordt $n \geq 20$ gebruikt.
[uit: *Statistiek om mee werken*; *Buijs(1999)]*

Het idee dat de kansen bij de binomiale verdeling berekend kunnen worden met gebruikmaking van een geschikt gekozen normale verdeling, is op te merken door de kansen van een binomiale verdeling (voor niet al te kleine *n*) in een grafiek te zetten. Uit de linker grafiek van figuur 11 valt te zien dat de vorm van de grafiek sterke gelijkenis vertoont met een grafiek van de normale verdeling.



Figuur 11: *Links: Bin(50,0.5) Rechts: Bin (50,0.05)*

Deze normale benadering mag alleen toegepast indien aan bepaalde voorwaarden is voldaan, namelijk indien het aantal pogingen $n \geq 20$. Bovendien moet ook gelden dat:

$$n\pi \geq 5$$

en

$$n(1-\pi) \geq 5$$

Indien dit niet *allebei* het geval is, mag de normale verdeling niet worden toegepast. Dit wordt zichtbaar indien je voor zo'n situatie een grafiek van de binomiale verdeling maakt. Dan blijkt dat de binomiale verdeling een nogal scheve vorm heeft, zodat op voorhand al te zien is dat daarbij geen mooie symmetrische normale verdeling kan worden toegepast. In de rechter grafiek van figuur 11 is dit voor $n=20$ en $\pi = 0.05$ te zien. Indien dus niet aan alle twee de voorwaarden wordt voldaan, wordt de *poisson benadering* toegepast.

Fracties

In nauwe relatie tot de binomiale verdeling staat het begrip *fractie*. De binomiale verdeling houdt zich bezig met het tellen van succeskansen bij een gegeven steekproefomvang *n*. Het aantal successen bedraagt dan hoogstens *n*: bij alle 'trekkingen' is er dan een succes geregistreerd. In de praktijk is het vaak nuttig om het aantal successen als een percentage of een fractie van het aantal pogingen weer te geven. Niet het absolute aantal successen wordt dan vermeld, maar het relatieve aantal wordt berekend. De fractie successen is het waargenomen aantal *(k)* gedeeld door het aantal pogingen *(n).*

Bij een gegeven succeskans π en een steekproefomvang *n* kan men spreken over de (onzekere) steekproeffractie $\underline{p}$.

De steekproeffractie $\underline{p}$ is een kansvariabele omdat het aantal successen $\underline{k}$ een kansvariabele is. De verdeling van $\underline{p}$ is dan ook rechtstreeks af te leiden uit de

verdeling van $\underline{k}$: $P(\underline{p} = \frac{\underline{k}}{n}) = P(\underline{k} = k)$. Omdat het berekenen van steekproeffracties vooral interessant is wanneer de steekproefomvang enigszins groot is, wordt het geval bekeken waarbij de normale benadering gebruikt wordt bij de binomiaal verdeelde variabele $\underline{k}$. Voor $\underline{k}$ geldt: E($\underline{k}$)=$n\pi$ en Var($\underline{k}$)=$n\pi$ *(1-* π*).* Dus:

$$E(\underline{p}) = E\left(\frac{\underline{k}}{n}\right) = \frac{1}{n} \cdot E(\underline{k}) = \frac{1}{n} \cdot n\pi = \pi$$

Hier uit kan je lezen dat bij een succeskans π (de populatiefractie) een steekproeffractie altijd de verwachtingswaarde π heeft, ongeacht de keuze van *n*. Verder geldt:

$$Var(\underline{p}) = Var(\frac{\underline{k}}{n}) = \frac{1}{n^2} Var(\underline{k}) = \frac{1}{n^2} n\pi(1-\pi) = \frac{\pi(1-\pi)}{n}$$

Dus $\sigma_{\underline{p}} = \sqrt{\frac{\pi(1-\pi)}{n}}$

Uit deze formule kan opgemerkt worden dat een grotere waarde van *n* leidt tot een kleinere waarde van $\sigma_{\underline{p}}$.

Betrouwbaarheidinterval fractie

De steekproeffractie $p = \frac{k}{n}$ geeft dus weer hoeveel successen er in een steekproef voorkomen. Deze fractie betekent niet dat in de populatie ook precies dezelfde p-waarde geldt. Het is dus van belang om grenzen te formuleren waartussen, de populatiefractie (of succeskans) zich vermoedelijk zal bevinden. Dit wordt het betrouwbaarheidsinterval voor een fractie genoemd.

Als eerst wordt bekeken hoe de kansverdeling van $\underline{p}$ (de waargenomen fractie) afhangt van π (de populatiefractie). Een veronderstelling is dat de populatie oneindig groot is. Hierdoor mag verondersteld worden dat bij elke trekking (met of zonder teruglegging) de succeskans gelijk is aan π. Het aantal successen $\underline{k}$ bij een steekproef van omvang *n* volgt dan een binomiale verdeling.
Hiervoor geldt:

$$E(\underline{k}) = n\pi$$
$$Var(\underline{k}) = n\pi(1-\pi)$$

Voor de steekproeffractie $\underline{p} = \frac{\underline{k}}{n}$ geldt dus:

$$E(\underline{p}) = \frac{n\pi}{n} = \pi$$
$$Var(\underline{p}) = \frac{n\pi(1-\pi)}{n^2} = \frac{\pi(1-\pi)}{n}$$

Wat aan deze formules valt te zien dat de onbekende waarde π ook in de variantieformule zit. Dit probleem, wordt afhankelijk van de steekproefomvang, op verschillende manieren aangepakt.

Bij het construeren van schattingsintervallen wordt er onderscheidt gemaakt tussen twee gevallen:

a) *n* is klein (n<200)
b) *n* is groot (n≥200)

ad a

Als *n < 200* dan wordt de normale benadering toegepast. Het doel is om een interval te construeren met een gegeven betrouwbaarheid *α*. Het resultaat wordt weergegeven door de volgende formule[5]:

$$\pi_{1,2} = \frac{p + \frac{z^2}{2n} \pm z\sqrt{\frac{p(1-p)}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \qquad (1)$$

De linkergrens π_1 volgt als in de formule bij het symbool het minteken wordt gebruikt. Met het plusteken krijgt men de rechtergrens π_2. Bij een gekozen betrouwbaarheid α volgt de *z-waarde*[6] uit de tabel van de normale verdeling. Het betrouwbaarheidsinterval luidt:

$$\pi_1 < \pi < \pi_2$$

ad b

Als *n ≥200*, dan wordt een eenvoudige variant toegepast van de formule *(1)* om een schattinginterval te geven van de populatiefractie. Nu wordt de in de steekproef gevonden waarde voor *p* gebruikt in plaats van de onbekende π bij de berekening van de standaarddeviatie. Het gevraagde interval luidt dan:

$$p - z\sqrt{\frac{p(1-p)}{n}} < \pi < p + z\sqrt{\frac{p(1-p)}{n}} \qquad (2)$$

Een andere manier om dit te zien is dat als we in formule (1) *n* groot laten worden dat deze formule vanzelf overgaat in formule (2) .

[5] Uit Statistiek om mee verder te werken ; Buijs(1999)
[6] Zie hoofdstuk 6.1.1 voor de definitie van *z-waarde* en *betrouwbaarheid*

Verschil toets fracties

Stel dat er twee steekproeven zijn en we onderzoeken of de populatiefracties (=succeskansen) π_1 en π_2 van de populaties waaruit de steekproeven zijn getrokken gelijk aan elkaar kunnen zijn. Hierbij wordt aangenomen dat de twee steekproeven mogen worden beschouwd als onderling onafhankelijke experimenten. Verder wordt aangenomen dat de steekproeven voldoende groot zijn om bij berekeningen de normale benadering toe te staan.

Stel dat in steekproef *1* $p_1 = \frac{k_1}{n_1}$ als waargenomen fractie gevonden wordt en voor steekproef *2*: $p2 = \frac{k_2}{n_2}$ geldt.

Voor de populatiefractie wordt de volgende hypothese geformuleerd:

$$H_0 : \pi_1 - \pi_2 = 0$$
$$H_1 : \pi_1 - \pi_2 \neq 0$$

Er wordt een significantieniveau *α* gekozen. In dit onderzoek wordt *α=0.05* genomen. Als toetsingsgrootheid wordt $\underline{d} = p_1 - p_2$ gebruikt. Hiervoor geldt onder aanname van de nulhypothese, wegens onafhankelijkheid:

$$E(\underline{d})=E(\underline{p}_1-\underline{p}_2)=E(\underline{p}_1)-E(\underline{p}_2)= \pi_1 - \pi_2 = 0$$

Verder geldt:

$$Var(\underline{p}_1) = \frac{\pi_1(1-\pi_1)}{n_1}$$
$$Var(\underline{p}_2) = \frac{\pi_2(1-\pi_2)}{n_2}$$

Dus wegens onderlinge onafhankelijkheid van de experimenten:

$$Var(\underline{d}) = Var(\underline{p}_1 - \underline{p}_2) = Var(p_1) + Var(p_2)$$
$$= \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$$

Helaas is deze variantie niet te berekenen omdat π_1 *en* π_2 niet gegeven zijn. Er kan echter gebruik worden gemaakt van de waargenomen fracties p_1 en p_2.

Er volgt dan: $\sigma_{\underline{d}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Met $\sigma_{\underline{d}}$ kunnen de grenzen van het kritieke gebied eenvoudig berekend worden bij de gekozen *α*. $\sigma_{\underline{d}}$ kan echter ook op een andere manier bepaald worden.

Bij een nulhypothese $H_0 : \pi_1 - \pi_2 = 0$ wordt er verondersteld dat beide populatiefracties aan elkaar gelijk zijn. De grootheden $p_1 = \frac{k_1}{n_1}$ en $p_2 = \frac{k_2}{n_2}$ zijn dan schattingen van dezelfde onbekende π. Deze twee grootheden kunnen gecombineerd worden tot een schatter *p** door het totale aantal successen $k_1 + k_2$ te delen door het totale aantal trekkingen $n_1 + n_2$.

Als schatter van π_1 en π_2 wordt dan $p^* = \frac{k_1 + k_2}{n_1 + n_2}$ gebruikt.

Voor $\sigma_{\underline{d}}$ geldt dan het volgende: $\sigma_{\underline{d}} = \sqrt{\frac{p^*(1-p^*)}{n_1} + \frac{p^*(1-p^*)}{n_2}}$

Oftewel: $\sigma_{\underline{d}} = \sqrt{p^*(1-p^*)(\frac{1}{n_1} + \frac{1}{n_2})}$

Als kritiek gebied wordt dan gevonden:

$Z=\{x|\ \ x< -z\sigma_{\underline{d}}\ \ of\ \ x> z\sigma_{\underline{d}}\}$

Hierbij is *z* opgezocht in de tabel van de normale verdeling. Als het geconstateerde verschil $p_1 - p_2$ tot het kritiek gebied *Z* behoort dan wordt de nulhypothese verworpen. Dit betekent dat het verschil significant is. Hierboven is de nulhypothese $H_0 : \pi_1 - \pi_2 = 0$ gebruikt. Bij een nulhypothese $H_0 : \pi_1 - \pi_2 = \Delta$ (met $\Delta \neq 0$) geldt een analoge opzet van de toets. Voor $\sigma_{\underline{d}}$ moet dan de formule:

$\sigma_{\underline{d}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ worden gebruikt.

Het kritieke gebied is dan:

$Z=\{x|\ \ x< \Delta - z\sigma_{\underline{d}}\ \ of\ x> \Delta + z\sigma_{\underline{d}}\}$

[uit: *Statistiek om mee te werken*; Buijs (1999)]

Deze methode is in het programma *Evaluatie Tool* van de Postbank geautomatiseerd.

6.1.3 Bootstrappen

Inleiding bootstrap

Bootstrap is een methode die gebruikt kan worden om een onbekende grootheid te benaderen.
Bij de Bootstrap is geen verdelingsaanname nodig. Het enige dat bij Bootstrap dient te worden aangenomen is dat de steekproef representatief is voor de gehele populatie. Zonder moderne computers zou de Bootstrap praktisch niet bruikbaar zijn. In plaats van het theoretisch afleiden van een kansverdeling wordt bij de Bootstrap namelijk gebruik gemaakt van computersimulaties. Het algemene idee bij Bootstrap is om aan de hand van de getrokken steekproef, steekproeven te simuleren. Omdat de steekproef waaruit getrokken wordt representatief is voor de gehele populatie (volgens bovenstaande aanname), kunnen we aan de hand van de *B* gesimuleerde steekproeven een uitspraak doen over de gehele populatie. Op basis van één steekproef, kan èn het gemiddelde van de populatie geschat worden, èn een inschatting gemaakt worden van de waarschijnlijke onnauwkeurigheid van die schatting. Voor dit laatste wordt de bootstrap procedure gebruikt. In praktijk komt het erop neer dat een bootstrap-steekproef verkregen wordt door uit de oorspronkelijke steekproef van grootte *n* een nieuwe steekproef van grootte *n* te trekken via *trekking met teruglegging*.
[uit: Dictaat Vrije Universiteit *Statistische Data Analyse*; Gunst M. de (2005)]

Simulaties in dit onderzoek

In dit onderzoek wordt door middel van simulaties, betrouwbaarheidsintervallen rondom het bruto/netto percentage en de netto hypotheek productie opgesteld. Deze simulaties worden voor zowel de actiegroep als de referentiegroep gedaan.
Als eerste wordt besproken hoe het bruto/netto percentages worden gesimuleerd en hoe hieruit een betrouwbaarheidsinterval volgt. Daarna wordt besproken hoe de bruto hypotheek productie wordt gesimuleerd en hoe hieruit een betrouwbaarheidsinterval volgt.

1.Het simuleren van het bruto/netto percentage.

Dit percentage is het percentage netto respons dat volgt uit de bruto respons. Dit kan als volgt geformuleerd worden:

$$\textbf{bruto/netto percentage} = \frac{aantal\ \ netto\ \ eerste\ \ contacten\ \ in\ \ de\ \ eerste\ \ n\ \ weken}{aantal\ \ bruto\ \ eerste\ \ contacten\ \ in\ \ de\ \ eerste\ \ n\ \ weken} \qquad \textbf{(1)}$$

Met:
Aantal netto eerste contacten=
[% (aantal eerste contacten actiegroep) - %(aantal eerste contacten referentiegroep)]
x grootte mailgroep.
Aantal bruto eerste contacten =aantal eerste contacten van de actiegroep.

Om tot dit percentage te komen moeten dus de aantallen eerste contacten in de eerste n weken worden gesimuleerd. Dit voor zowel de actiegroep als de referentiegroep. Deze aantallen kunnen gesimuleerd worden door middel van een kans. Hiermee definieer je dus een kansverdeling. Bij de bootstrap procedure wordt

aangenomen dat de verdeling van de aantallen eerste contacten in de eerste n weken in de steekproef identiek is aan die van de populatie. Deze genoemde kans is de kans dat een klant een eerste contact in de eerste n weken heeft. Deze kans is gelijk aan:

$$P(\text{klant heeft eerste contact in de eerste n weken}) = \frac{\text{totaal aantal klanten met eerste contact in de eerste n weken}}{\text{grootte groep}}$$

Deze kans kan zowel voor de actiegroep als referentiegroep berekend worden: bij *grootte groep* moet dan resp. de grootte van de actiegroep of grootte van de referentiegroep staan. Aangezien elk van de personen in de actiegroep (stel deze is 125.000 groot) door 'een willekeurig persoon' vervangen kan worden, kan een nieuwe steekproef van 125.000 personen gemaakt worden maken met aantallen klanten die een eerste contacten hebben in de eerste n weken, die willekeurig getrokken zijn uit de aantallen klanten genoemd met de bijbehorende kansverdeling. Deze nieuwe steekproef heet dan de 'bootstrap-steekproef'. Deze procedure wordt dan 100 herhaald om 100 van zulke bootstrap-steekproeven te krijgen. Dit voor zowel de actiegroep als de referentiegroep.

Na 100 simulaties krijg je voor elke steekproef een aantal eerste contacten in de eerste n weken voor zowel de actiegroep als de referentiegroep. Met behulp van formule *(1)* kan vervolgens voor elke steekproef het bruto/netto percentage bepaald worden. Bij 100 keer simuleren volgen dus 100 bruto/netto percentages.

95% betrouwbaarheidsinterval rondom bruto/netto percentage

Om een 95% betrouwbaarheidsinterval te nemen sorteer je de waardes van de 100 bruto/netto percentages verkregen na 100 keer te simuleren. De waardes die buiten het 95% interval vallen zijn dus de 3 (=2.5% *100) kleinste en de 3 grootste waardes. Zo ontstaat dus een betrouwbaarheidsinterval voor het bruto/netto percentage.

2. Het simuleren van bruto productie

Om een betrouwbaarheidsinterval rondom de netto productie te krijgen wordt als eerste de bruto hypotheek productie die volgt uit de bruto eerste contacten van de eerste n weken gesimuleerd. Deze aantallen kunnen ook weer gesimuleerd worden door middel van een kans. Voor de actiegroep kan namelijk bepaald worden wat de kans is dat een klant, die in de eerste n weken een eerste contact heeft, een hypotheek afsluit. Deze kans is namelijk:

$$P(\text{klant met ec in eerste n weken sluit een hypotheek}) = \frac{\text{aantal klanten met ec in de eerste n weken die een hypotheek afsluiten}}{\text{totaal aantal klanten met eerste contact in de eerste n weken}}$$

Door middel van deze kans kan de bruto productie die volgt uit de contacten van de eerste n weken gesimuleerd worden. Na 100 simulaties wordt voor elke steekproef een bruto productie berekend. Bij 100 simulaties volgen dus 100 bruto productie aantallen.

Om de netto hypotheek productie te krijgen wordt voor elke steekproef de verkregen bruto hypotheek productie vermenigvuldigd met het bruto/netto percentage uit de vorige simulatie:

netto hypotheek productie simulatie i =

[bruto hypotheek productie simulatie i] x [bruto/netto percentage simulatie i]

95% betrouwbaarheidsinterval rondom /netto hypotheek productie

Om een 95% betrouwbaarheidsinterval rondom de netto productie te nemen sorteer je de waardes van de 100 netto hypotheekproducties verkregen na 100 keer te simuleren. De waardes die buiten het 95% interval vallen zijn dus de 3 (=2.5% *100) kleinste en de 3 grootste waardes. Zo ontstaat dus een 95% betrouwbaarheidsinterval voor de netto hypotheekproductie.

De VBA code die gebruikt is om deze genoemde aantallen te simuleren in Excel, is te zien in output 2. In deze code worden de hierboven besproken simulaties uitgevoerd.

```
Sub bootstrap()
   For i = 1 To Range(100)
     Cells(1 + i, 10) = 0
     Cells(1 + i, 11) = 0
     Cells(1 + i, 12) = 0

     For j = 1 To Range("grootte mailing actiegroep")
       If Rnd < Range("Pec_b") Then
            Cells(1 + i, 10) = Cells(1 + i, 10) + 1
            If Rnd < Range("p_conv") Then
               Cells(1 + i, 11) = Cells(1 + i, 11) + 1
            End If
       End If
     Next
     For j = 1 To Range("grootte mailing referentiegroep")
       If Rnd < Range("Pec_r") Then
            Cells(1 + i, 12) = Cells(1 + i, 12) + 1

       End If
     Next
   Next
End Sub
```

Output 2: Simulatie bruto eerste contacten en hypotheken actiegroep en refgroep

6.2 Actie productie

6.3 Conclusie

Op basis van dit deel, kan worden geconcludeerd worden dat de eerste n weken van de eerste contacten een goede benadering geven van de hypotheekproductie die getriggered is door de actie.

7 Onderzoek Deel 1: Tijdreeks Analyse

7.1 Inleiding

Een tijdreeks is een verzameling waarnemingen die sequentieel door de tijd gemaakt zijn. Tijdreeks analyse bevat methoden die dergelijke tijdreeksen proberen te begrijpen, vaak om of de onderliggende theorie van de data punten te begrijpen (Waar komen ze vandaan? Wat heeft ze gegenereerd?) of voor het maken van voorspellingen. De onderzoeksvraag voor dit deel van het onderzoek is:

Kan het responsverloop van de eerste contacten door middel van tijdreeks modellen gemodelleerd worden?

Er worden een aantal verschillende notaties gebruikt voor tijdreeksen analyse:

$$Y = \{Y_1, Y_2, \ldots\}$$

is een vaak gebruikte notatie voor het specificeren van een tijdreeks.

In de volgende paragrafen wordt als eerste de theorie van tijdreeksen besproken. Daarna wordt de tijdreeks van de eerste contacten geanalyseerd. Meer aanvullende informatie over dit onderwerp kan in *The Analysis of Time Series* (Chatfield C.) gevonden worden.

7.2 Tijdplot

De eerste en meest belangrijke stap in de tijdreeks analyse, is het plotten van de observaties tegen de tijd. Deze grafiek, ook wel *tijdplot* genoemd, kan belangrijke waarnemingen zoals trend, seizoenspatronen, uitschieters en discontinuïteiten weergeven.
Voor het modelleren van tijdreeksen is het van belang dat een reeks stationair is en dus geen trends of seizoenspatronen vertoont. Omdat het vaak in de praktijk voor komt dat een tijdreeks trends of seizoenspatronen vertoont, is het vaak noodzakelijk deze niet-stationaire reeks te transformeren in een stationaire reeks. Hoe dit gedaan wordt, wordt in de volgende paragraaf besproken.

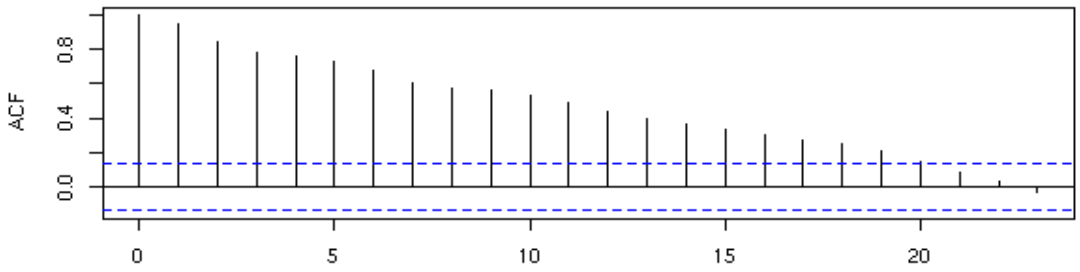
7.3 Stationaire reeksen

Zojuist is het begrip stationaire reeks genoemd. Een tijdreeks is stationair als voor de variabele Y_t geldt:

$$E(Y_t) = \mu$$
$$Var(Y_t) = \sigma^2$$

Deze eis moet voor elk tijdstip *t* gelden, wat aangeeft dat de gemiddelde waarde en de standaarddeviatie onafhankelijk zijn van de tijd. In het algemeen betekent dit dat de tijdreeks geen trend bevat.

De standaard manier om te weten of een tijdreeks niet-stationair is door het plotten van de reeks en zijn autocorrelatiecoëfficiënt[7]. Door de grafiek van de reeks over de tijd bestuderen kan onderzocht worden of er een zichtbare trend is. Als de reeks niet-stationair is zal de autocorrelatie functie vaak langzaam afnemen (zie figuur 11).



Figuur 11: *Autocorrelogram van een niet- stationaire reeks*

Als een reeks niet stationair is dan moet deze met behulp van bepaalde technieken stationair worden gemaakt voordat een model hierop gebouwd kan worden.
Dit houdt in dat de trends verwijderd moeten worden uit het gemiddelde en de variantie van de reeks.

Als het gemiddelde een trend bevat, betekent dit dat de gemiddelde waarde van de reeks stijgt of daalt in de tijd. Als de variantie een trend bevat dan zal variabiliteit van de reeks stijgen of dalen in de tijd. Hoe kunnen deze verwijderd worden uit de reeks? Er zijn verschillende methodes om dit te doen.

<u>Stabiliseren van de variantie</u>
Om de variantie te stabiliseren van een reeks, kan de Box-Cox Power transformatie toegepast worden. Gegeven een tijdreeks $\{y_t\}$ en een transformatie parameter λ, dan wordt de transformatie gegeven door:

$$y_t = \frac{(y_t^{\lambda} - 1)}{\lambda} \text{ if } \lambda \neq 0$$

$$y_t = \log(y_t) \text{ if } \lambda = 0$$

In het algemeen, is de LOG transformatie (λ=0) een goede keuze om de toenemende variantie in een reeks te verwijderen.

[7] Het begrip *autocorrelatie* wordt in paragraaf 7.4 besproken.

Differencing
De overgebleven trend kan dan verwijderd worden door toepassing van eerste verschillen ("differencing"). De analyse met autocorrelaties wordt dan niet toegepast op de oorspronkelijke tijdreeks Y_t maar op de reeks ΔY_t. Deze nieuwe reeks is Y_{t-t-1}. Dit proces kan herhaald worden totdat de trend verdwenen is.

7.3.1 Toets voor stationariteit

In de literatuur zijn verschillende toetsten bekend die gebruikt kunnen worden om te bepalen of een reeks stationair is of niet. Een bekende toets is de *Philips-Perron unit root* toets. Deze toets is gebaseerd op autoregressieve modellen. De vergelijking van een autoregressieve AR(1)-model is als volgt gedefinieerd:

$$X_t = \alpha\ X_{t-1} + Z_t \qquad (1)$$

waarbij Z_t het witte ruis component is en α het richtingscoëfficiënt.
Een unit root is aanwezig als het volgende geldt: α =1. Als een tijdreeks een unit root bevat dan bevat de tijdreeks een stochastische trend is dus niet stationair. Dit volgt uit het volgende feit dat:
[uit: *Econometric Models & Economic Forecasts* (Pindyck (1991)]

$$X_{t-1} = \alpha X_{t-2} + Z_{t-1} \qquad (2)$$

$$X_{t-2} = \alpha X_{t-3} + Z_{t-2} \qquad (3)$$

Als *(2)* gesubstitueerd wordt in *(1)* volgt:

$$X_t = \alpha(\alpha X_{t-2} + Z_{t-1}) + Z_t = \alpha^2 X_{t-2} + \alpha Z_{t-1} + Z_t \qquad (4)$$

Als *(3)* gesubstitueerd wordt in *(4)* volgt:

$$X_t = \alpha^2(\alpha X_{t-3} + Z_{t-2}) + \alpha Z_{t-1} + Z_t = \alpha^3 X_{t-3} + \alpha^2 Z_{t-1} + \alpha Z_{t-1} + Z_t \qquad (5)$$

Opeenvolgende substituties van deze soort leiden tot:

$$X_t = \alpha^t X_0 + \alpha\ Z_{t-1} + \alpha^2 Z_{t-2} + \alpha^3 Z_{t-3} + \ldots \alpha^t Z_0 + Z_t \qquad (6)$$

Er worden 3 gevallen onderscheiden:

- $\alpha < 1 \Rightarrow \alpha^t \to 0 \quad als \quad t \to \infty$

  De trillingen in het systeem dempen geleidelijk uit.

- $\alpha = 1 \Rightarrow \alpha^t = 1 \quad \forall \quad t$

  De trillingen blijven in het systeem en dempen niet uit. Je krijgt dan:

$$X_t = X_0 + \sum_{i=0}^{\infty} Z_t \qquad als \quad t \to \infty$$

- $\alpha > 1$. De trillingen in het systeem krijgen meer invloed voor grote *t* omdat als $\alpha > 1$ dan $\alpha^3 > \alpha^2 > \alpha$ etc.

Bij de *Philips-Perron unit root* toets is de nul hypothese als volgt:

H_0: α = 1 (reeks is niet stationair)

H_1: $|\alpha|$< 1 (reeks is stationair)

Als betrouwbaarheid wordt *α=0.05* genomen.

Meer informatie over *de Philips-Perron* unit root is te vinden in *Econometric Models & Economic Forecasts* (Pindyck e.a. 1991).

7.3.2 Toets voor witte ruis

In de literatuur zijn verschillende toetsten voor witte ruis bekend. Een bekende toets is *de Box-Ljung* toets. Het idee hierachter is het bepalen van de (gewogen) som van de eerste autocorrelatiecoëfficiënten. Deze som volgt bij benadering een Chi^2 verdeling. De *Ljung-Box* is een variant van de *Box-Pierce* en geeft een betere Chi^2 benadering voor kleine steekproeven.

De *Ljung-Box* test is gebaseerd op de autocorrelatie plotjes. In plaats van het toetsten van witte ruis op elke willekeurige lag, toetst het de “overall” randomness gebaseerd op een aantal lags. De *Ljung-Box* toets kan als volgt gedefinieerd worden:

H_0 : De data is random (witte ruis).

H_1 : De data is niet random (geen witte ruis).

De toetsinggrootheid is:

$$Q_{LB} = T(T+2)\sum_{j=1}^{M} \frac{r_j^2}{T-j}$$

met *T* de grootte van de steekproef, r_j de autocorrelatie at lag *j*, en *M* is het aantal lags dat gebruikt wordt . Als betrouwbaarheid wordt *α=0.05* genomen.

De witte ruis hypothese wordt verworpen als $Q_{LB} > \mathrm{X}^2_{1-\alpha,M}$ met $\mathrm{X}^2_{1-\alpha,M}$ Chi-kwadraat verdeeld bij *M* vrijheidsgraden.

7.4 Autocorrelatie

Een heel belangrijk onderdeel bij het analyseren van tijdreeksen betreft het bestuderen van de onderlinge afhankelijkheid van opeenvolgende waarnemingen Y_t in zo'n reeks. Hierbij kan het zijn dat een waarneming Y_t een verband toont met de waarneming van één periode eerder, maar het kan ook zijn dat waarnemingen die verder weg liggen een invloed op Y_t hebben. Een eenvoudige manier om dit te onderzoeken, is het toepassen van autocorrelatie.

Hierbij wordt in beginsel van dezelfde formule gebruik gemaakt als bij gewone correlatie, waar de grootheid *r* de mate van lineaire samenhang van de variabelen *X* en *Y* aangaf. Nu wordt er gewerkt met de variabelen Y_t en Y_{t-1}. De 'getallenparen' zijn nu $(Y_2, Y_1), (Y_3, Y_2)$ enz.

Wanneer de autocorrelatie wordt berekend van een variabele op twee opeenvolgende tijdstippen, spreekt men van een vertragingsfactor één (*lag* = 1). Uiteraard kan ook autocorrelatie voorkomen tussen waarden die meer dan één tijdseenheid uit elkaar liggen. Zo is de autocorrelatie voor ieder gewenst tijdsinterval te berekenen (*lag* = *2, 3, . . . , k*).

De gebruikelijke formule voor de autocorrelatiecoëfficiënt luidt:

$$r_1 = \frac{\sum_{t=2}^{n}(Y_t - \bar{Y})(Y_{t-1} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} = \frac{\sum_{t=2}^{n} y_t y_{t-1}}{\sum_{t=1}^{n} y_t^2}$$

De coëfficiënt r_1 wordt de autocorrelatiecoëfficiënt met time-lag *1* genoemd. Op een soortgelijke manier kan de autocorrelatiecoëfficiënt met een time-lag van *k* perioden gedefinieerd worden. In formule:

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} = \frac{\sum_{t=k+1}^{n} y_t y_{t-k}}{\sum_{t=1}^{n} y_t^2}$$

De autocorrelatiecoëfficiënten die berekend zijn op basis van beschikbare (historische) gegevens vormen een belangrijk hulpmiddel bij het opsporen van bepaalde patronen in een tijdreeks.
Het uitzetten van de autocorrelatie tegen het aantal *lags* in een grafiek of tabel biedt een belangrijk instrument voor de beoordeling van tijdreeksen. Dit wordt de autocorrelatie functie (ACF) genoemd .

De ACF geeft ook aanwijzingen voor het al dan niet stationair zijn van een tijdreeks. Als vuistregel wordt gehanteerd dat autocorrelaties tussen $2/N^{1/2}$ en $-2/N^{1/2}$ (waarbij N gelijk is aan het aantal waarnemingen) *niet statistisch* significant zijn.

Vaak worden berekende autocorrelatiecoëfficiënten aangegeven in een speciale grafische voorstelling: het *autocorrelogram*. Het belang van zo'n autocorrelogram is dat hiermee op het gebied van tijdreeks analyse vrij snel onderkend kan worden wat voor onderliggende structuur deze reeks heeft.

Een reeks met een trend of met een seizoenspatroon zal een autocorrelogram voortbrengen dat een zeer herkenbare vorm heeft. Een belangrijk thema bij tijdreeks analyse is echter of bij een stationaire reeks, dat is een reeks waarin trend-en seizoeninvloeden verwijderd zijn, nog steeds bepaalde afhankelijkheden bij opeenvolgende waarnemingen geconstateerd kunnen worden door middel van autocorrelatiecoëfficiënten.

7.4.1 Interpretatie autocorrelogram

Een handig hulpmiddel bij het interpreteren van autocorrelatiecoëfficiënten is de autocorrelogram waar de autocorrelatiecoëfficiënten r_k geplot worden tegen lag k for $k=0,1,\ldots,M$ waarbij M veel kleiner dan de groep waarnemingen is.

Het interpreteren van een verzameling autocorrelatiecoëfficiënten is geen gemakkelijke zaak. Hier volgen enkele richtlijnen.

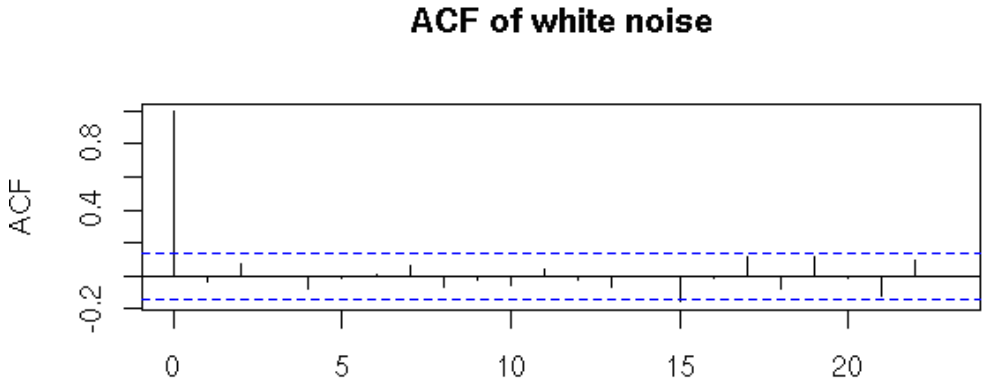
Random Reeks
Een tijdreeks is random als het bestaat uit een reeks random variabelen , $\{Y_t\}$, die onafhankelijk van elkaar zijn en identiek verdeeld. Vaak wordt aangenomen dat random variabelen normaal verdeeld zijn met μ=0 and variance σ^2_Y. Uit de definitie volgt dan dat het proces een constante μ en variantie heeft. De onafhankelijkheid aanname houdt in dat:

$$Cov(Y_t\, Y_{t+k}) = \begin{cases} \sigma_Y^2 & k = 0 \\ 0 & k = \pm 1, \pm 2, \ldots \end{cases}$$

Dit betekent dat de verschillende waardes ongecorreleerd zijn zodat de autocorrelatiecoëfficiënt gegeven wordt door:

$$\rho_k = \begin{cases} 1 & k = 0 \\ 0 & k = \pm 1, \pm 2, \ldots \end{cases}$$

Dus voor grote N, is de verwachting dat $r_k \approx 0$ voor alle niet-0 waarden van k. In figuur 12 is een autocorrelogram van witte ruis te zien.



Figuur 12: *Autocorrelogram van witte ruis*

7.4.2 Toets voor autocorrelatie

Evenals bij 'gewone' correlatiecoëfficiënten wordt bij autocorrelatie een onderscheid gemaakt tussen populatie- en steekproefwaarden. Het idee is hierbij dat een verzamelde reeks gegevens ten gevolge van steekproeftoeval een waarde voor een autocorrelatiecoëfficiënt r_k kan tonen die iets afwijkt van een populatiecoëfficiënt ρ_k. Zo zal bij een populatiecoëfficiënt $\rho_k=0$ niet elke verzamelde reeks Y_t-waarden een r_k tonen die exact gelijk aan 0 is. Om te onderzoeken of een r_k significant van 0 afwijkt, is een toets ontworpen die als volgt is opgezet.

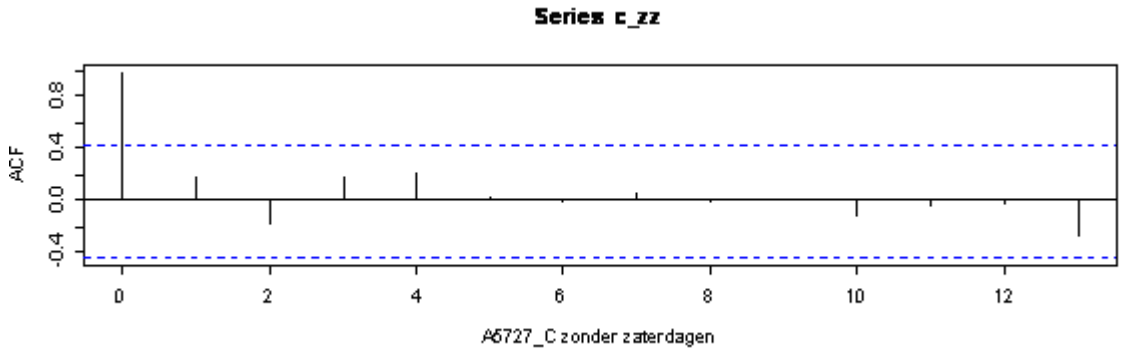
De hypothesen zijn:

$$H_0 : \rho_k = 0$$
$$H_1 : \rho_k \neq 0$$

De grootheid r_k is de toetsinggrootheid waarvoor aangetoond is dat deze bij benadering een normale verdeling volgt met als standaarddeviatie $1/\sqrt{n}$. Meestal wordt deze toets uitgevoerd met $\alpha=0.05$, waarbij gewerkt wordt met $z=2$ (eigenlijk z=1.96). Als een berekende autocorrelatiecoëfficiënt vervolgens buiten het gebied $(-2/\sqrt{n}\ ,\ 2/\sqrt{n})$ valt, dan wordt H_0 verworpen.

Voorbeeld
In de figuur 13 is de autocorrelogram te zien van actie A5727_C zonder de respons op de zaterdagen. De twee blauwe lijnen geven het gebied aan waar H_0 verworpen wordt. Deze tijdreeks heeft 20 data punten, dus de lijnen zijn ter hoogte van $\pm\frac{2}{\sqrt{20}} = \pm 0.44$. Alle waardes behalve at lag 0 vallen binnen de grenzen. De nul hypothese kan in alle gevallen verworpen worden. Dit heeft als gevolg dat de autocorrelatiecoëfficiënten allemaal niet significant van 0 verschillen. Deze datapunten zijn dus ongecorreleerd.

De autocorrelograms van de andere acties zijn in bijlage C terug te vinden.



Figuur 13: *Autocorrelogram actie A5727_C*

7.5 Tijdreeks onderzoek

De besproken theorie wordt nu toegepast op de drie reeksen van de eerste contacten. Als eerste wordt onderzocht of de reeksen stationair zijn en vervolgens worden ze op witte ruis getoetst.

Stationariteit

Als eerste wordt elke actie op stationariteit onderzocht . Het statistische programma R heeft een functie genaamd `PP.test` die de reeks op stationariteit toetst m.b.v *Philips-Perron unit root test*. Deze toets wordt voor elke actie uitgevoerd. De reeks wordt zowel met de zaterdagen getoetst als zonder. Dit omdat de respons op zaterdagen heel laag is. De p-waardes zijn in tabel 19 per actie samengevat. De output van R is te zien in de bijlage C.

Tabel 19

P-waarde stationariteit

Actie	*P-waarde (data met zaterdagen)*	*P-waarde (data zonder zaterdagen)*
A6144	*0.01*	*0.01*
A5727_B	*0.01768*	*0.1733*
A5727_C	*0.02690*	*0.01*

Uit tabel 19 valt te zien dat in alle gevallen behalve 1, de nulhypothese dat de reeks niet stationair is wordt verworpen bij een betrouwbaarheid van $\alpha=0.05$. De reeksen zijn dus stationair. Dit in tegenstelling dus tot actie *A5727_B zonder de respons op zaterdagen*, waarbij de nulhypothese aangenomen moet worden. Dit betekent dat deze reeks niet stationair is. Het nemen van de eerste difference bij deze reeks is een manier om stationariteit te krijgen. Dit kan in `R` gemakkelijk gedaan worden door de functie `diff`.

Na de *Philips-Perron* unit root test op deze gedifferenced data van actie A5727_B toe te passen, wordt bij een gevonden p-waarde van 0.01, de nul hypothese verworpen. Deze reeks is nu stationair.

Witte ruis

Nu dat elke reeks stationair is, kan de data op witte ruis worden getoetst . Dit kan gedaan worden door de *Box-Ljung* toets toe te passen. In `R` wordt de functie `Box.test(data , type="Ljung-Box")` hiervoor gebruikt.
De resultaten van deze toets zijn samengevat in tabel 20.

Tabel 20

P-waarde

Actie	*P-waarde (data met zaterdagen)*	*P-waarde (data zonder zaterdagen)*
A6144	*0.3402*	*0.01933*
A5727_B	*0.2257*	*0.3703*
A5727_C	*0.2254*	*0.3843*

Uit deze tabel volgt dat de nulhypothese dat de data witte ruis is, in alle gevallen behalve één, niet verworpen kan worden bij een betrouwbaarheid van 0.05. De reeksen van actie A5727_B en actie A5727_C zijn dus witte ruis processen.

7.6 Conclusie

Het responsverloop van de eerste kan contacten voor twee acties gezien worden als een witte ruis proces. Een witte ruis proces is een reeks van ongecorreleerde random variabelen, waarvan de verwachting en de variantie constant is. Met andere woorden, de waargenomen respons aantallen per dag zijn identiek verdeelde onafhankelijke random variabelen. In het algemeen kan het responsverloop dus niet gemodelleerd worden een tijdreeks proces.

8 Onderzoek deel II: Classificeren van klanten

8.1 Inleiding

Als klanten geclassificeerd zouden kunnen worden op basis van hun responstijd[8] dan zou je voor een groep nieuwe klanten kunnen voorspellen hoe snel en wanneer ze responderen. De onderzoeksvraag voor dit deel van het onderzoek is:

Zijn er klanten met bepaalde eigenschappen die altijd vroeg of laat responderen?

Om dit onderzoek te kunnen doen worden als eerste verschillende variabelen aan elke klant toegekend. Daarna worden deze variabelen onderzocht door middel van statistische technieken op lineaire verbanden met de responstijd. Op basis van deze variabelen wordt dus geprobeerd de responstijd van een klant te voorspellen.

8.2 Toekennen klant variabelen

Aan elke klant worden verschillende variabelen uit de databases *Kozmos, Geomarkt* en *Acxiom* gekoppeld. Dit zijn databases die verschillende variabelen van klanten bevatten. Hieronder wordt beschreven wat voor soort klant kenmerken elke database bevat.

KOZMOS

Kozmos staat voor *Klant aggregaat voor OnderZoek, MOdelbouw en Scoring.* In Kozmos zijn onder andere gegevens terug te vinden over:

- *Klantkenmerken*: kenmerken van de persoon zelf.
- *Bezitkenmerken:* het wel of niet bezitten van producten, deze variabelen beginnen met een B. Bij alle bezitsvariabelen bezit de klant het product als de klant ultimo maand minimaal een lopende overeenkomst van het product heeft.
- *Saldomeetwaarden*: saldi per klant voor producten of groepen van producten, deze variabelen beginnen met een S.
- *Transactiemeetwaarden*: totale aantallen transacties per maand, deze variabelen beginnen met een T.

Geomarkt

Sinds 1985 verzamelt Wegener DM consumentengegevens met als doel de kennis van klanten en prospects te vergroten. Deze GeoMarktprofiel klantkennis geeft antwoord op de volgende vragen:

- Wie zijn mijn klanten en prospects (demografie en geografie): welstand, opleiding, levensfase, geografische herkomst, etc.
- Wat consumeren mijn klanten en prospects (consumptiegedrag): mediagedrag, winkelgedrag, vrijetijdsbesteding, etc.
- Waarom consumeren mijn klanten en prospects (psychografie): risicobereidheid, sociale betrokkenheid, impulsiviteit, etc.

[8] Met responstijd wordt de tijd tussen de mailing en een eerste contact bedoeld.

Acxiom
InfoBase is een database die op adresniveau een aantal kenmerken bewaart, zoals gezinssamenstelling, inkomensniveau, opleidingniveau en eigendomsindicatie van het woonhuis. Deels wordt de data 1-op-1 opgeleverd uit de CCI consumentenenquête dat jaarlijks door een half miljoen consumenten wordt ingevuld. Deels bestaat de InfoBase uit inschattingen van deze kenmerken. Deze inschattingen worden mede gemaakt met behulp van klantendatabases van zakenpartners van Acxiom. Dit zijn bv. postorderbedrijven, uitgeverijen, krantenconcerns en een keur van andere partners uit diverse branches.

8.3 Classificeren

Classificatie technieken hebben als doel om te voorspellen tot welke groep (klasse) een bepaald object behoort. De objecten worden beschreven door een verzameling variabelen. Een *object* is bijvoorbeeld een klant. De klant wordt beschreven door een aantal kenmerken, zoals bijvoorbeeld leeftijd, geslacht, saldo betaalrekening. Stel dat men op basis van deze kenmerken wil voorspellen wanneer de klant gaat reageren op een bepaalde mail. De kenmerken 'Leeftijd,' 'Geslacht' en 'Saldo betaalrekening' zijn de *onafhankelijke* oftewel *verklarende variabelen* en de variabele 'Responstijd' is de *afhankelijke variabele* oftewel de *klasse* waar de klant toe behoort. De rijen in tabel 21 stellen de objecten oftewel klanten voor en de kolommen zijn de variabelen.

Tabel 21

Klant variabelen

Leeftijd	***Geslacht***	***Saldo betaalrekening***	***Responstijd (in dagen)***
21	*M*	*550*	*3*
28	*V*	*300*	*5*
27	*M*	*210*	*11*

Training en Test set
Voor het maken van een model is er data nodig met objecten waarvan de klassen bekend zijn. Deze dataset wordt gesplitst in twee datasets, namelijk de training set en de test set. De verhouding is meestal 75% voor de training set en 25% voor de test set. In de training set zijn zowel de waarden van de verklarende variabelen bekend als de waarde van de klassen van de verschillende objecten. De bedoeling is om classificatieregels te vinden, zodat nieuwe objecten geclassificeerd kunnen worden. Vervolgens kan het model getest worden op de *test set*; door middel van de classificatieregels worden de objecten van de test set geclassificeerd. Dit resultaat kan dan vergeleken worden met de werkelijke klassen behorende bij de objecten van de test set. De test set wordt dus gebruikt om de algemeenheid van de classificatieregels te testen; gelden de classificatieregels ook voor nieuwe data? Op deze manier kan de validiteit van het model bepaald worden. Bij de Postbank wordt de *liftcurve* gebruikt om de voorspelkracht van het model te bepalen.

8.4 Lineaire regressie

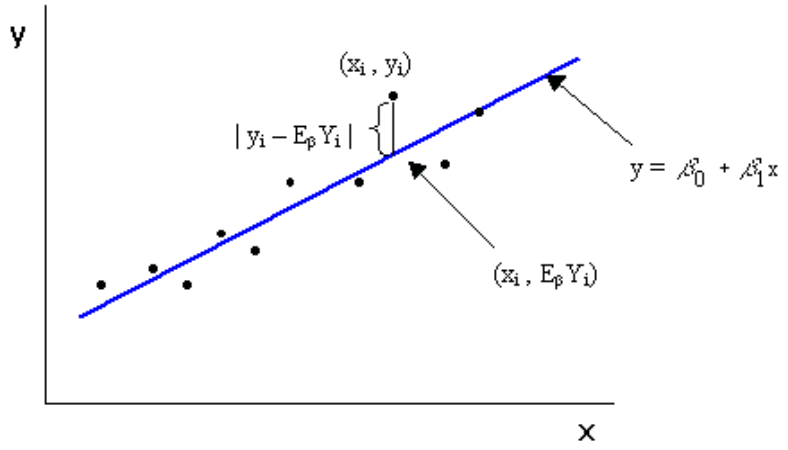
Een statistische methode om te onderzoeken hoe de *responstijd* variabele afhangt van een of meerdere of meerdere variabele is *regressie*. De variabele die men wenst te begrijpen, te verklaren of te voorspellen noemt men de *afhankelijke variabele* of de *respons variabele*. De variabelen die in het model worden opgenomen omdat er vermoed wordt dat deze een invloed kunnen hebben op de afhankelijke variabele, noemt men de *onafhankelijke variabele* of de *verklarende variabele*. Er wordt gesproken van *enkelvoudige regressie* als de responstijd maar van één verklarende variabele afhangt, en van *meervoudige regressie* als de responstijd van meerdere verklarende variabele afhangt. In dit onderzoek is de afhankelijke variabele de *responstijd* en de onafhankelijke variabelen *alle* variabelen uit Kosmos, Acxiom en Geomarkt.

Bij lineaire regressie wordt de relatie tussen de betrokken variabelen weergegeven door een lineaire vergelijking. Hieronder wordt dit verder besproken.

Bij lineaire regressie bestaat het volgende verband tussen de onafhankelijke variabele(n) en de afhankelijke variabele:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} + \varepsilon_i \qquad \text{voor } i = 1, 2, \ldots, n,$$

met n het aantal observaties. De parameter β_0 wordt ook wel het *intercept* genoemd en is de verwachte waarde van y_i wanneer alle onafhankelijke variabelen gelijk zijn aan nul. De parameters $\beta_1, \ldots, \beta_m$ heten de *regressiecoëfficiënten* en kunnen geïnterpreteerd worden als de verandering in de verwachte waarde van y_i als de bijbehorende onafhankelijke variabele met één eenheid stijgt én de overige onafhankelijke variabelen constant gehouden worden. De parameter ε_i staat voor de *fout* behorende bij observatie *i*. De fout ε_i is de verticale afstand tussen de observatie y_i en het punt op de regressiecurve behorende bij die observatie (zie figuur 14). Er wordt verondersteld dat de fouttermen onafhankelijk en normaal verdeeld zijn.



Figuur 14: *Voorbeeld Lineaire Regressie in twee dimensionale ruimte*

In geval van twee variabelen, de afhankelijke variabele en één onafhankelijke variabele, wordt de regressiefunctie beschreven door een rechte lijn (zie figuur 14). Hier spreekt men van een enkelvoudige lineaire regressie. Indien er meerdere variabelen aanwezig zijn, spreekt men van meervoudige lineaire regressie en wordt de regressiefunctie beschreven door een hypervlak (= de uitbreiding van het begrip 'rechte' voor hogere dimensies).

Sommige niet-lineaire verbanden kunnen ook gemodelleerd worden door middel van de lineaire regressie techniek. De vergelijking $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_m x^m + \varepsilon$ bijvoorbeeld, kan omgezet worden naar het lineaire model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \varepsilon$ waarbij $x_1 = x$, $x_2 = x^2$, …, $x_m = x^m$.

Parameterschatting

In het geval van lineaire regressie zijn er in de literatuur enkele methodes bekend om de waarden van de parameters $\beta_0, \ldots, \beta_m$ te bepalen. De meest gebruikelijke methode is de kleinste kwadraten methode.

Kleinste kwadratenprincipe

Beschouw vergelijking 3.1. Laat $E_\beta Y_i$ de verwachte waarde van de observatie y_i zijn bij parameterkeuze β. De functie,

$$S(\beta) = \sum_{i=1}^{n} (y_i - E_\beta Y_i)^2 \qquad (3.2)$$

meet de som van de kwadraten van de afstanden tussen de observaties y_i en de punten van de kromme (E_β y_i) met overeenkomstige x-coördinaten, oftewel het verschil tussen de werkelijke waarde van de observaties en de verwachte waarden wanneer β de echte parameter zou zijn. Een voorbeeld hiervan is weergeven in Figuur 10. De parametervector β kan geschat worden door de vector $\hat{\beta}$ die de functie $S(\beta)$ minimaliseert. Deze methode staat ook wel bekend als de *kleinste kwadratenmethode*. De gevonden parametervector $\hat{\beta}$ bepaalt de vergelijking van de beste kromme.

8.5 Kwaliteit van het model

Liftcurve

In de marketing wordt veelal de liftcurve gebruikt om de voorspel kracht van een model te bepalen. In het geval van een continue target variabele, zoals bij dit onderzoek, is de interpretatie van een liftcurve anders dan bij het geval van een binaire target variabele. Nu zijn er namelijk geen niet-respondenten. Iedereen heeft namelijk al gerespondeerd. Voor nieuwe klanten die gaan responderen, moet bepaald worden *wanneer* ze gaan responderen.

In de grafiek van de liftcurve staat nu op de x-as het cumulatieve totaal aantal klanten (de respondenten). Op de y-as staat vervolgens het cumulatieve aantal

responsdagen. Deze kan op basis van de data van de training set bepaald worden. Voor deze set is namelijk bekend wanneer elke klant gerespondeerd heeft. Om een liftcurve voor een continue target variabele te kunnen interpreteren wordt als eerste de optimale liftcurve gemaakt op basis van de werkelijke datagegevens: voor elke klant is namelijk de responstijd bekend.

Dit wordt voor actie A6144 besproken. De klanten worden als eerste gesorteerd in afnemende grootte op basis van de responstijd. Op de horizontale as komen dan de verschillende selectiegroottes als percentage van het totaal aantal klanten (hier: x) te staan. De verticale as geeft het aantal cumulatieve responsdagen weer van het totale aantal klanten (hier: y). Deze optimale liftcurve is weergegeven in grafiek 10 met daarin ook de verkregen liftcurve van de training set en test set.

Een optimale liftcurve is ook aan de grafiek toegevoegd. Deze liftcurve is voor een binaire target variabele gemakkelijk uit te leggen. Stel dat de target variabele *responderen* is. Deze variabele kan alleen de waardes *ja* of *nee* aannemen, dus of een klant op een actie respondeert of niet. De ideale situatie is dan dat bij een kleine selectie van klanten (10%) een zo hoog mogelijk responspercentage (100%) wordt behaald. Als voor elke klantselectie, het responspercentage 100% is, heb je een perfect model. Voor een continue target variabele kan de optimale liftcurve gemaakt worden op basis van de werkelijke datagegevens, zoals eerder al beschreven.

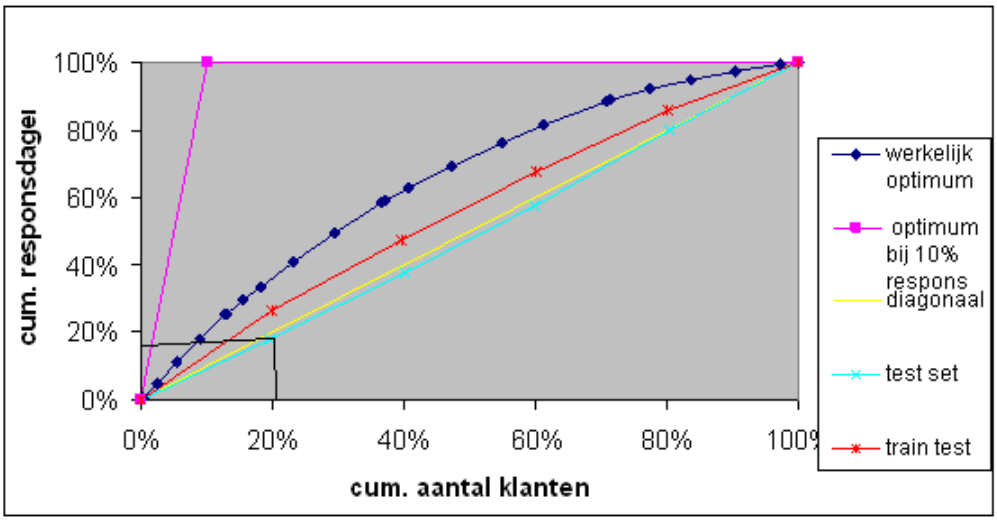
Uit de liftcurven van grafiek 10 is te zien dat er 5 klassen van klanten zijn gemaakt die elk een eigen gemiddelde responstijd hebben. Deze responstijd kan als volgt uit de grafiek worden gehaald: de richtingscoëfficiënt (RC) van elk klasse geeft de gemiddelde responstijd van die klantengroep weer.

RC per klasse =gemiddelde responstijd per klantgroep

Interpretatie

In grafiek is het punt (19.63%, 18.15%) gemarkeerd. In aantallen terug gerekend is dit punt gelijk aan (x , y).

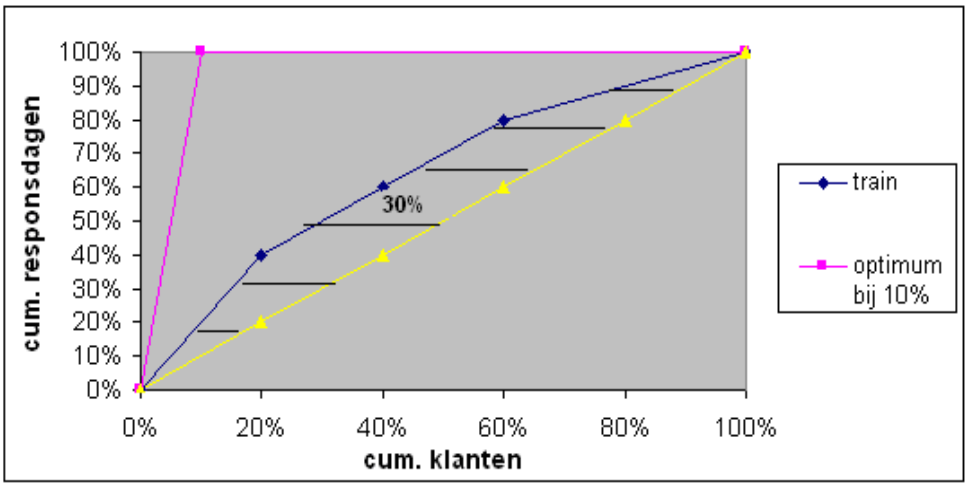
Uit grafiek 10 is te zien, door de bijna rechte lijn van zowel de training set als test set, dat de verschillende klassen onderling weinig van elkaar verschillen in responstijd.



Grafiek 10: Liftcurve actie A6144

C-waarde

De Postbank gebruikt een *c-waarde* om de kwaliteit van een liftcurve te bepalen. Deze c-waarde is de oppervlakte onder de liftcurve. De ideale situatie is een c-waarde van 1. Dit gebeurt alleen bij het optimum. In praktijk is deze c-waarde niet realistisch. De Postbank heeft daarom een ondergrens voor de kwaliteit van het model vastgesteld: de behaalde liftcurve moet minstens 30% van de oppervlakte boven de diagonaal bevatten (zie grafiek 11). De hierbij horende c-waarde is dan 0.65 (=0.5 + 30%*0.5). Dus een model met een c-waarde lager dan 0.65 wordt beschouwd als een slecht model.



Grafiek 11: Ondergrens liftcurve

Determinatiecoëfficiënt

De *determinatiecoëfficiënt* is een maat voor het deel van de variatie in de waarnemingen, dat door het lineaire regressiemodel wordt verklaard. Als eerste wordt de spreiding gedefinieerd:

Totale spreiding	=	*Verklaarde Spreiding*	+	*Onverklaarde spreiding*
SST	=	*SSR*	+	*SSE*
$\sum (Y_i - \bar{Y})^2$	=	$\sum (\hat{Y}_i - \bar{Y})^2$	+	$\sum (\hat{Y}_i - Y_i)^2$

De *determinatiecoëfficiënt* is gedefinieerd als:

$$R^2 = \frac{SST - SSE}{SST} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

waarbij:

- *SST* := de totale kwadraatsom.
- *SSE* := de gekwadrateerde verschillen tussen de werkelijk gemeten waarden Y_i en de aan de hand van het model voorspelde waarden $\hat{Y}_i$.
- *n* := het aantal observaties.
- $\bar{Y}_i$:= het gemiddelde van de geobserveerde Y_i waarden.

SST geeft de fout weer wanneer het gemiddelde van de observaties als voorspelling van de waarnemingen genomen wordt. SSE geeft de fout weer indien voor de voorspelling van de waarnemingen het regressiemodel gebruikt wordt.
R^2 geeft dus het percentage van de reductie van de fouten weer indien, in plaats van het gemiddelde van de waarnemingen, het regressiemodel gebruikt wordt. De waarde van R^2 ligt steeds tussen 0 en 1. Hoe dichter R^2 bij 1 ligt, hoe beter de werkelijke waarden van de afhankelijke variabele benaderd worden door het model. Als R^2 gelijk is aan 0, dan wil dit zeggen dat het model geen enkele toegevoegde waarde heeft. Het model past dan even goed als het model waarin geen enkele verklarende variabele voorkomt.

8.6 Onderzoek naar meervoudige verbanden

Bij lineaire regressie bestaat het volgende verband tussen meerdere onafhankelijke variabelen en de afhankelijke variabele *responstijd*:

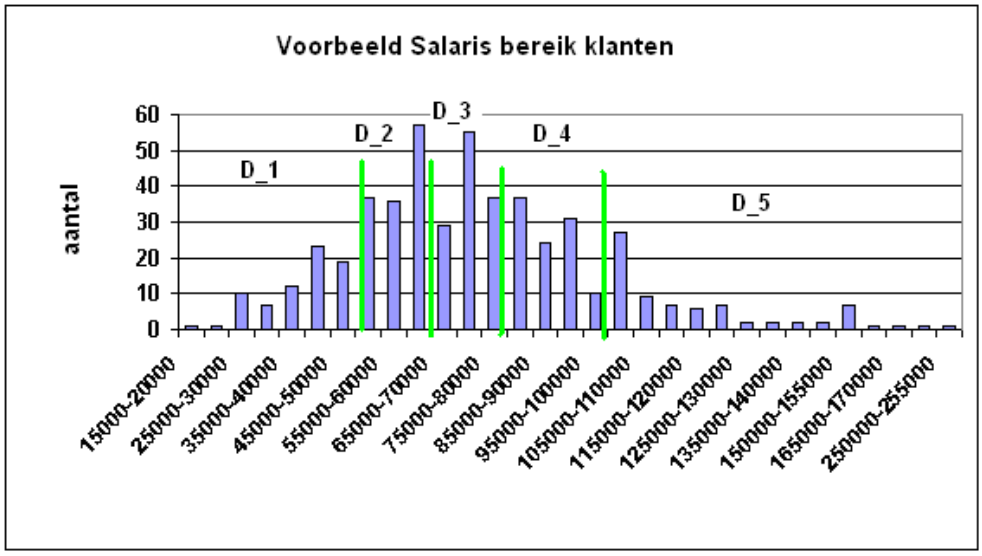
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} + \varepsilon_i \qquad \text{voor } i = 1, 2, \ldots, n,$$

Bij de Postbank is bekend dat de meeste variabelen geen onderlinge lineaire verbanden hebben. Ook is hier bekend dat de Postbank variabelen veel last van uitschieters hebben. Een aanname bij regressie is juist dat er een lineair verband tussen de afhankelijk en onafhankelijke variabelen is.

Dummies
Een oplossing voor deze Postbank variabelen is om deze in stukken op te hakken (dummies). Daarna wordt d.m.v. de *t-toets* getoetst of de gemiddelde waarde van de afhankelijke variabele in de verschillende dummies significant van elkaar verschillen. Als ze niet significant zijn, dan worden de dummies samengevoegd. Uiteindelijk worden de dummies gebruikt om een nieuwe variabele te maken die voor elke

dummieklasse de waarde heeft van het gemiddelde van de afhankelijke variabele in die dummieklasse. Hierdoor krijgt de nieuwe variabele een lineair verband met de afhankelijke variabele.

Voorbeeld
In grafiek 12 is een histogram gemaakt van de variabele *Salaris*. Hier is te zien dat deze variabele een grote spreiding heeft waarbij steeds minder mensen een hoog salaris hebben. De variabele wordt in 5 dummies opgehakt. In elke dummie kan een gemiddelde waarde voor de afhankelijke variabele worden bepaald.



Grafiek 12: *Dummies maken van de variabele Salaris*

Voor elke dummie D_1,D_2, D_3, D_4, D_5 is deze waarde bijvoorbeeld resp. 5,6,9,3,12 responsdagen. Met behulp van de t-toets wordt bepaald of de waardes van de afhankelijke variabele in de verschillende dummies significant verschillen. Stel dat de waardes allemaal significant van elkaar verschillen dan wordt er een nieuwe variabele genaamd *nieuwe_variabele* die de volgende waardes kan aannemen {3 ,5,6,9,12}. Hierdoor krijgt deze nieuwe variabele een lineair verband met de afhankelijke variabele.

Nu kan het meervoudige regressie model gebouwd worden. De methode die hiervoor gebruikt wordt heet *stapsgewijze regressie*. Stapsgewijze regressie voegt alle variabelen 'stapsgewijs' toe totdat de meest significante variabelen overblijven. Deze methode wordt hieronder verder beschreven.

Stapsgewijze regressie:
Met behulp van stapsgewijze regressie kan bepaald worden welke variabelen wel en niet in de regressievergelijking thuishoren. De volgende stappen worden genomen:

- Begin met 1 variabele en kijkt of de p-waarde[9] kleiner is dan α.
  Zo niet, dan wordt de variabele niet opgenomen.

- Zo wel, dan wordt de variabele opgenomen . Een tweede variabele wordt dan toegevoegd, waarvan de p-waarde met α vergeleken wordt.

[9] Zie hoofdstuk 6.1.1 voor de definitie van *p-waarde*.

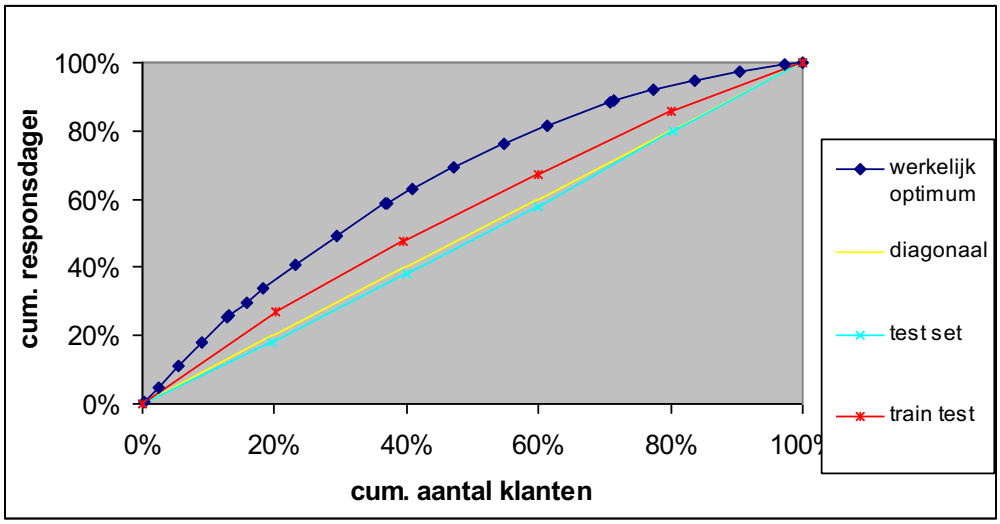
Is deze kleiner dan α, dan wordt ook deze opgenomen en wordt een derde variabele toegevoegd.

Zo gaat dit door tot er geen variabelen meer gevonden kunnen worden met een p-waarde kleiner dan α. Bij elke stap wordt steeds de variabele met de kleinste p-waarde toegevoegd. De variabele die al in het model zitten worden verwijderd als ze niet significant worden na het toevoegen van andere variabelen.

De output van de stapsgewijze regressie is in bijlage F terug te vinden.

Voorspelkracht Model

De voorspelkracht van het model kan in termen van liftcurves worden weergegeven. In grafiek 13 is de liftcurve behorende bij actie A6144 te zien. Uit deze grafiek is te zien, door de bijna rechte lijn van zowel de training set als test set, dat de verschillende klassen onderling weinig van elkaar verschillen in responstijd. Het model maakt dus nauwelijks onderscheid tussen de verschillende klassen. Deze liftcurves duiden dus op een bijzonder slecht model. Dit geldt ook voor de liftcurves van actie A5727_B en actie A5727_C. Deze liftcurves zijn de bijlage E terug te vinden.



Grafiek 13: *Liftcurve van actie A6144*

Kwaliteit Model

De kwaliteit van het model kan in termen van de determinatiecoëfficiënt worden weergegeven. In tabel 22 zijn de determinatiecoëfficiënten te zien van de verkregen modellen voor elke actie. Het verkregen coëfficiënt van R^2=0,0636 voor actie A6144 willen zeggen, dat een klein deel van de variantie, namelijk. 6,36%, door het lineaire regressiemodel wordt verklaard. Dit is dus een bijzonder slecht model. Hetzelfde geldt ook voor de andere twee regressiemodellen.

Tabel 22

Determinatiecoëfficiënten

Actie	***R^2***
A6144	*0,0636*
A5727_B	*0,1536*
A5727_C	*0,0944*

8.7 Conclusie

In paragraaf 8.6 is de variabele responstijd onderzocht op meervoudige verbanden. De gevonden liftcurves en slechte determinatiecoëfficiënten duiden erop dat er geen meervoudige verbanden zijn. Er kan dus geconcludeerd worden dat de variabele *responstijd* ***geen verband*** heeft met de beschikbare variabelen. De *responstijd* van een klant is dus een variabele die niet voorspeld kan worden op basis van de beschikbare klant kenmerken.

9 Onderzoek deel III: Modelbouw door Poisson benadering

9.1 Inleiding

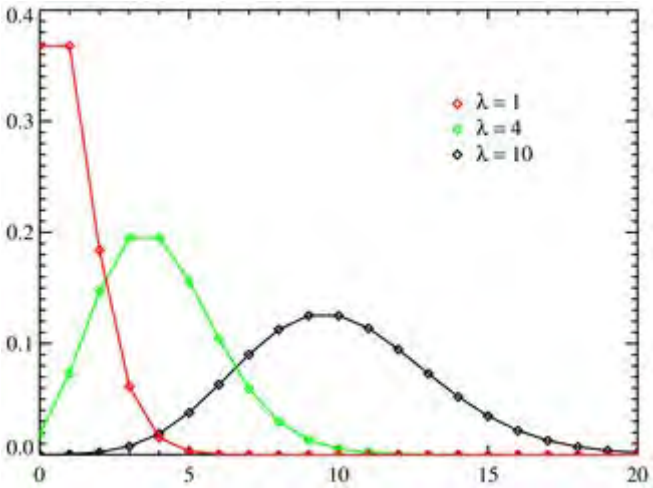
In het Tijdreeks Analyse onderzoek is geconstateerd dat de respons per dag gedurende de n weken een stationair *witte ruis* proces is. De aantallen respondenten per dag kunnen dus onderling onafhankelijk en identiek verdeeld worden verondersteld. Als het aantal van de gehele actiegroep heel groot wordt verondersteld en vervolgens de succeskans dat een persoon in de actiegroep respondeert klein en de personen onafhankelijk van elkaar responderen dan kan het proces van het binnenkomen van de respons als een homogeen poisson proces worden gemodelleerd.

Elke actie heeft een eigen parameter λ die voor elke actie geschat wordt op basis van de data van die actie alleen. Dit kun je na 1 dag doen, na 2 dagen, na 3 dagen etc. Naar verwachting wordt steeds beter, als het model redelijk is. Na elke dag kan je dan de parameter van het poisson proces beter schatten en dus steeds beter de totale respons voorspellen.

9.2 Poisson verdeling

De poisson verdeling heeft net als de binomiale verdeling betrekking op het tellen van successen. Bij een binomiale verdeling is de steekproef omvang *n* gegeven, en is de succeskans π bekend (en bij iedere poging hetzelfde). De poisson verdeling kan voor grote *N* gebruikt worden bij met modelleren van het aantal successen. De parameter λ is dan het verwachte aantal successen.

Een belangrijke eigenschap van de poisson verdeling is dat de som van twee poisson variabelen weer een variabele is met een poisson verdeling met als parameter de som van de twee parameters.



Figuur 13: *Poisson verdeling voor λ=1,4,10*

Gegeven is een tijdsinterval van lengte *t* en een parameter λ (die het gemiddelde of verwachte aantal successen per tijdseenheid aangeeft).Voor de kans op een aantal successen *k* geldt dan de volgende formule:

$$P(\underline{k} = k) = \frac{(\lambda t)^k}{k!} \cdot e^{-\lambda t} \quad \textit{voor k=0,1,2}$$

In plaats van (λ *t)* wordt vaak het symbool *μ* gebruikt.

$$P(\underline{k} = k) = \frac{(\mu)^k}{k!} \cdot e^{-\mu} \quad \textit{voor k=0,1,2}$$

De verwachtingwaarde en variantie van $\underline{k}$:

$$E(\underline{k}) = \mu$$
$$Var(\underline{k}) = \mu$$

Benadering met behulp van normale verdeling

Indien *μ* groot is ($\mu \geq 10$), dan wordt de normale verdeling toegepast om de kansen te berekenen. Een kansvariabele $\underline{k}$ die poisson verdeeld is met parameter $\mu \geq 10$, kan benaderd worden door een variabele $\underline{x}$ die normaal verdeeld is.

$\underline{k} \sim Poisson(\mu)$ wordt dan benaderd door $\underline{x} \sim N(\mu, \sqrt{\mu})$

Toepassing bij de binomiale verdeling

De poisson verdeling kan in een bepaald speciaal geval gebruikt worden om kansen te berekenen voor een variabele die eigenlijk binomiaal verdeeld is. Het gaan dan om gevallen waarbij het aantal pogingen *n* altijd zeer groot is en de succeskans π ofwel zeer klein is (bijna 0) ofwel zeer groot is (bijna 1). Meer hierover is in hoofdstuk 6.1.2 besproken.

Het homogene Poisson proces

Wegens de stationariteit aanname dat volgt uit het tijdreeks onderzoek, is er wellicht sprake van een homogeen poisson proces. Hier volgt de definitie van een homogeen poisson proces:

Het proces *N(t)* op $[0,\infty)$ is een homogeen poisson proces met parameter λ als:

- *N(s, t)* een poisson verdeling heeft met verwachting $\lambda(t - s)$ voor alle $0 \leq s < t$;
- *N(s, t)* en $N(s_0, t_0)$ stochastisch onafhankelijk zijn voor alle $0 \leq s < t \leq s_0 < t_0$.

*[*uit*: Optimization of Business Processes;G.Koole]*

9.3 Schatting parameter

Definieer met *N(t)* het aantal responsen t/m dag t bij keuze tijdseenheid 1 dag.
Stel dat we een homogeen poisson proces hebben in [0,∞]met een onbekende parameter λ. Dan is $\frac{N(t)}{t}$ een goede schatter voor λ. Dit omdat $N(t) = \sum_{s=1}^{t} N(s-1,s)$ met alle $N(s-1,s)$ onafhankelijk en identiek verdeeld is .
De *wet van grote aantallen* zegt dan dat gemiddelde van een rij van stochastische variabelen met eenzelfde kansverdeling convergeert naar de gemeenschappelijke verwachtingswaarde, indien de grootte van de rij naar oneindig gaat. Dus als de wet van de grote aantallen wordt toegepast dan $\frac{N(t)}{t} \to \lambda$.
*[*uit*: Optimization of Business Processes;G.Koole]*

<u>Zuivere schatter</u>
Een schatter T heet zuiver voor het schatten van *g(θ)* als *E(T) = g(θ)* voor alle *θ*.
Een schatter is dus een zuivere schatter als de verwachting van de schatter gelijk is aan de te schatten waarde.
Het proces $\frac{N(t)}{t}$ is een zuivere schatter voor λ omdat:
$E(\frac{N(t)}{t}) = \frac{1}{t} E(N(t))$ en wegens de definitie van een homogeen proces volgt dan: $\frac{1}{t} E(N(t)) = \frac{1}{t} \cdot \lambda \cdot t = \lambda$

Meer informatie over dit onderwerp is te vinden in het dictaat *Optimization of Business Processes van G.Koole* en het dictaat *Algemene statistiek* .

Definieer met $\hat{\lambda}(n)$ de geschatte λ op basis van de respons aantallen na *n* dagen. Een schatting voor het verwachte respons aantal per dag is $\hat{\lambda}(n)$. Een schatting voor λ na 1 dag is dan $\hat{\lambda}(1)$. Een schatting voor het verwacht respons aantal na 24 dagen is $24\,\hat{\lambda}(1)$.

9.4 Intervallen

In de literatuur zijn er twee soorten intervallen bekend, namelijk:
- *Betrouwbaarheidsintervallen*
- *Voorspellingsintervallen*

<u>Betrouwbaarheidsinterval</u>
In dit onderzoek is de echte parameter λ van de Poisson verdeling niet bekend. Als er dan echter waarnemingen bekend zijn (uit een actie), kan deze onbekende parameter wordt geschat met behulp van $\frac{N(t)}{t}$. De vraag is echter; hoe nauwkeurig

is de schatting van de parameter? Het verschil tussen de werkelijke waarde en de geschatte waarde geeft de mate van nauwkeurigheid aan van de schatting. Het 95% betrouwbaarheidsinterval geeft dan een 95% interval voor de onbekende parameter λ.

Voorspellingsinterval
Dit interval heeft betrekking tot het voorspellingsprobleem. Een voorspellingsprobleem is als een uitspraak over een afzonderlijke uitkomst gedaan wil worden. Omdat nu alleen naar een specifieke waarde wordt gekeken, moet niet alleen gekeken worden naar de onzekerheid in de schatting van de verwachting maar moet ook rekening worden gehouden met de afwijking van het aantal van zijn verwachting.
[uit *Statistiek om mee te werken, Buijs (1999)].*

Een opvallend verschil tussen de plots met betrouwbaarheidsintervallen en voorspellingsintervallen is in breedte van de intervallen. De breedte van de voorspellingsintervallen is breder omdat deze dus aan twee soorten onzekerheid onderhevig is. Deze intervallen worden hieronder nog afzonderlijk verder besproken.

9.4.1 Betrouwbaarheidsinterval

Het betrouwbaarheidsinterval geeft het gebied van waarden aan, waarbinnen zich de te schatten parameter kan bevinden. Meestal wordt een waarschijnlijkheid van 95% gebruikt. Dit betekent dat wanneer we het onderzoek zouden herhalen 95 van de 100 herhalingen een resultaat geven dat binnen het interval ligt. Het betrouwbaarheidsinterval zegt iets over de nauwkeurigheid van de berekende waarden.

Voor een dataset $x_1, x_2, \ldots, x_n$ die $N(\mu, \sigma^2)$ verdeeld is met bekende σ, wordt het 95% betrouwbaarheidinterval voor μ gegeven door:

$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

waarbij $z_{\alpha/2}$ de rechter kritieke waarde is van de N(0, 1) verdeling.

Om een 95% betrouwbaarheidsinterval rondom de λ op te stellen, kan de normale benadering toegepast worden. Deze benadering kan gebruikt worden omdat $\mu \geq 20$ is. Na deze benadering toegepast te hebben zoals in paragraaf 9.2 *Benadering met behulp van normale verdeling* is beschreven, volgt het volgende:

- het 95% betrouwbaarheidsinterval voor de **onbekende parameter λ**:

$$\hat{\lambda}(n) - 1{,}96 \cdot \frac{\sqrt{\hat{\lambda}(n)}}{\sqrt{n}} < \lambda < \hat{\lambda}(n) + 1{,}96 \cdot \frac{\sqrt{\hat{\lambda}(n)}}{\sqrt{n}}$$

- het 95% betrouwbaarheidsinterval voor de **verwachting van de totale respons** in de eerste maand *μ :*

$$24 \cdot \hat{\lambda}(n) - 1{,}96 \cdot 24 \cdot \frac{\sqrt{\hat{\lambda}(n)}}{\sqrt{n}} < \mu < 24 \cdot \hat{\lambda}(n) + 1{,}96 \cdot 24 \cdot \frac{\sqrt{\hat{\lambda}(n)}}{\sqrt{n}} \qquad \textbf{(1)}$$

9.4.2 Voorspellingsinterval

Een voorspellingsinterval is een interval voor het voorspelde aantal. Een 95% voorspellingsinterval geeft het 95% interval van de voorspelde waarde. Een betrouwbaarheidsinterval wordt gebruikt voor het bepalen van interval schattingen voor de onbekende parameter. Een voorspellingsinterval is een interval schatting van een (onbekende) toekomstige waarde.

Als eerste wordt de theorie van voorwaardelijke verwachtingen en variantie besproken die voor het opzetten van het voorspellingsinterval nodig is.

Verwachting voorwaardelijke variantie

De *verwachting* en *variantie* van een kansvariabele hebben de volgende eigenschappen:

$$E(X + c) = E(X) + c$$
$$E(aX) = aE(X)$$
$$Var(aX + b) = a^2 Var(X)$$

De definitie van variantie is als volgt:

$$Var(X) = E(X^2) - (E(X))^2 \qquad \textbf{(2)}$$

De voorwaardelijke variantie is dan:

$$Var(X \mid Y) = E(X^2 \mid Y) + (E(X \mid Y))^2$$

en de verwachting van de voorwaardelijke variantie is dan:

$$E(Var(X \mid Y)) = E(E(X^2 \mid Y)) + E[(E(X \mid Y))^2]$$

Omdat $E(X) = E(E(X \mid Y))$ volgt:

$$E(Var(X \mid Y)) = E(X^2) + E[(E(X \mid Y))^2] \qquad \textbf{(3)}$$

Variantie van de voorwaardelijke verwachting

Door gebruik te maken van de formule van variantie (zie formule 2) volgt:

$$Var(E(X \mid Y)) = E([E(X \mid Y)]^2) - [E(E(X \mid Y))]^2$$

Omdat $E(X) = E(E(X \mid Y))$, volgt:

$$Var(E(X \mid Y)) = E([E(X \mid Y)]^2) - [E(X)]^2 \qquad \textbf{(4)}$$

Wat opvalt, is dat formule *(3)* and *(4)* beide de term $E([E(X \mid Y)]^2)$ bevatten, maar met tegengestelde tekens. Door deze twee vergelijkingen op te tellen volgt:

$$E(Var(X \mid Y)) + Var(E(X \mid Y)) = E(X^2) - [E(X)]^2$$

Het rechterdeel van deze vergelijking is gewoon $Var(X)$ (zie formule 2).Met andere woorden:

$$Var(X) = E(Var(X \mid Y)) + Var(E(X \mid Y)) \qquad \textbf{(5)}$$

Deze formule wordt straks gebruikt voor het opstellen van het voorspellingsinterval.

[Uit: *Probability and Random Processes* ; Grimmett G. Stirzaker D. (1992)]

Voorspeller

Het doel is om de totale respons te voorspellen. De totale respons (*k*) is een poisson variabele met parameter 24λ. Dus *k~Poisson(24* λ*).*

Na *n* dagen is de voorspeller (dit is een stochastische variabele) van de totale respons een poisson variabele met parameter $24\hat{\lambda}(n)$. Dus k_1*~Poisson(*$24\hat{\lambda}(n)$).

De voorspelling van de totale respons is dan de verwachting van deze variabele. Deze verwachting is, omdat $\hat{\lambda}(n)(= \frac{N(n)}{n})$ een zuivere schatter is voor λ:

$E(k_1) = 24\lambda$, maar de λ is nu juist de onbekende.

Als schatter hiervan neem je daarom niet de onvoorwaardelijke verwachting 24λ maar de voorwaardelijke verwachting gegeven de respons op de eerste n dagen:

$$E(k_1 \mid respons\ op\ de\ eerste\ n\ dagen) = 24\hat{\lambda}(n)$$

Zolang je de respons op de eerste *n* dagen niet weet dan, is dit ook een stochastische variabele met verwachting 24λ:

$$E(E(k_1 \mid respons\ op\ de\ eerste\ n\ dagen)) = 24\lambda$$

Zodra de waarden van de *n* responsen bekend zijn, dan kan de schatting $24\,\hat{\lambda}(n)$ van de voorspelling berekend worden.

Voor een exact voorspellingsinterval voor de totale respons moet de verdeling van de voorspeller bekend zijn. Omdat de parameter van de voorspeller (k_1~*Poisson(*$24\,\hat{\lambda}(n)$) een stochastische variabele is (is afhankelijk van de data), is deze verdeling nog niet bekend. Het enige wat bekend is, is dat deze voorwaardelijk gegeven de responsen op de eerste *n* dagen een poisson verdeling heeft. Dit omdat gegeven de responsen de parameter niet stochastisch is en dan volgens definitie van een homogeen poisson proces, poisson verdeeld is.

Hoewel de (onvoorwaardelijke)verdeling van de voorspeller niet bekend is, kan wel iets over de standaardafwijking van deze verdeling gezegd worden.

Standaardafwijking

De standaardafwijking van de voorspeller kan, door de theorie van voorwaardelijke kansen, door de volgende formule bepaald worden (zie formule (5)):

$$Var(k_1) = Var(E(k_1 \mid respons\ op\ de\ eerste\ n\ dagen)) + E(Var(k_1 \mid respons\ op\ de\ eerste\ n\ dagen))$$

Deze berekening wordt nu per som deel verder uitgelegd.

Eerste deel som

Als eerste wordt $Var(E(k_1 \mid respons\ op\ de\ eerste\ n\ dagen))$ berekend. Het binnenste deel, $E(k_1 \mid respons\ op\ de\ eerste\ n\ dagen)$, is gelijk aan $24\,\hat{\lambda}(n)$, zoals eerder al verteld. Nu moet dus $Var\,(24\,\hat{\lambda}(n))$ bepaald worden.

De $\hat{\lambda}(n)$ is bepaald door $\frac{N(n)}{n}$. Uit de definitie van een homogeen poisson proces, is bekend dat $N(n)$ poisson verdeeld is met verwachting λ **n* dus

$$Var(24\hat{\lambda}(n)) = Var(\frac{24}{n} \cdot n \cdot \hat{\lambda}(n)) = (\frac{24}{n})^2 Var(n \cdot \frac{N(n)}{n}) = (\frac{24}{n})^2 Var(N(n)) = (\frac{24}{n})^2 \cdot n \cdot \lambda = \frac{24^2}{n} \cdot \lambda$$

Twee deel som

Nu moet $E(Var(k_1 \mid respons\ op\ de\ eerste\ n\ dagen))$ berekend worden. Het binnenste deel $Var(k_1 \mid respons\ op\ de\ eerste\ n\ dagen)$ is ook weer gelijk aan $24\,\hat{\lambda}(n)$. Dit omdat bij de poisson verdeling de verwachting gelijk aan de variantie is (zie hoofdstuk 9.2).

Nu moet $E(24\hat{\lambda}(n))$ bepaald worden. Deze is gelijk aan $E(24\hat{\lambda}(n)) = 24\lambda$ omdat $\hat{\lambda}(n)$ een zuivere schatter is.

Uit deze twee delen volgt dus dat: $Var(k_1) = \frac{24^2}{n}\lambda + 24\lambda$

Dan is de standaardafwijking gelijk aan:

$$\sigma(k_1) = \sqrt{\frac{24^2}{n}\lambda + 24\lambda}$$

Hierin is de λ onbekend en die kan weer door $\hat{\lambda}(n)$ geschat worden:

$$\hat{\sigma}(k_1) = \sqrt{\frac{24^2}{n}\hat{\lambda}(n) + 24\hat{\lambda}(n)}$$

Een voorspellinginterval met ongeveer 95% betrouwbaarheid voor de totale respons tot en met dag 24 wordt dan gegeven door:

$$[\ 24\,\hat{\lambda}(n) - 2\times\hat{\sigma}(k_1)\ ,\ 24\,\hat{\lambda}(n) - 2\times\hat{\sigma}(k_1)\ ] \qquad \textbf{(6)}$$

Voorbeeld

Stel dat op voor een actie op dag 1, 27 respons gemeten is. Op basis van deze eerste dag, kan λ geschat worden. Deze schatting is dan: $\hat{\lambda}(1) = \frac{27}{1} = 27$. De voorspelde respons na 24 dagen is dan 27*24= 648. Om nu een voorspellingsinterval op te stellen, wordt formule *(6)* gebruikt:

$[\ 24\,\hat{\lambda}(n) - 2\times\hat{\sigma}(k_1)\ ,\ 24\,\hat{\lambda}(n) - 2\times\hat{\sigma}(k_1)\ ]$

De standaarddeviatie is: $\hat{\sigma}(k_1) = \sqrt{\frac{24^2}{n}\hat{\lambda}(n) + 24\hat{\lambda}(n)} = 127.28$

Het volgende voorspellingsinterval wordt dan verkregen:

[24*27 – 2*127.28 , 24*27+2*127.28]= [393.44 902.6].

Dus met 95% betrouwbaarheid zit de voorspelde respons in dit interval.

9.5 Modelbouw

De bovenstaande theorie wordt nu toegepast op de beschikbare acties. Elke actie heeft een eigen parameter λ die voor elke actie geschat wordt op basis van de data van die actie alleen. Dit kun je na 1 dag doen, na 2 dagen, na 3 dagen etc. Na elke dag kan je de parameter van het poisson proces beter schatten en dus steeds beter voorspellen. Daarna wordt om elke voorspelling een voorspellingsinterval opgesteld.

9.5.1 Model op actie A6144

Voor actie A6144 wordt de parameter λ geschat op basis van de data die elke dag binnenkomt. Daarna wordt de totale respons voorspeld na *n* dagen. Voor elke voorspelling wordt ook het voorspellingsinterval opgesteld. In de volgende paragrafen worden deze procedures besproken.

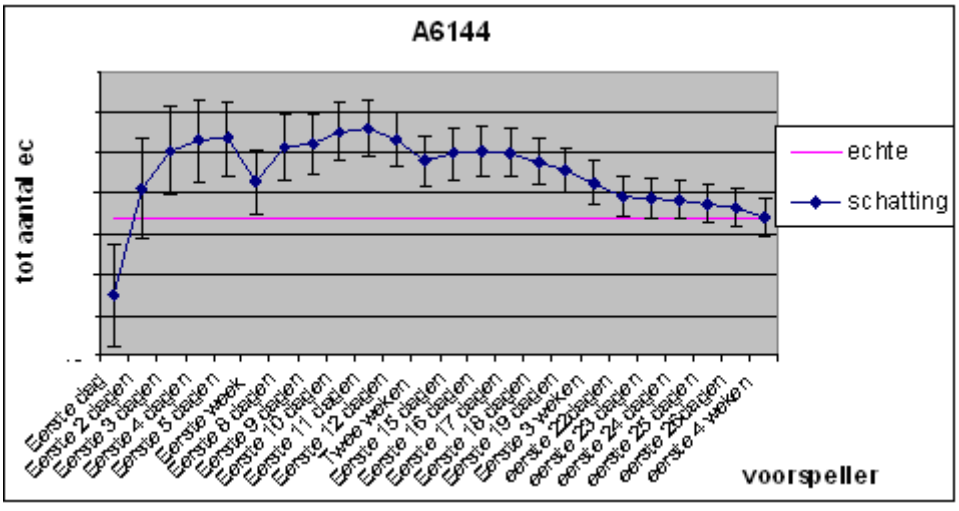
9.5.1.1 Schatten parameter λ

De parameter λ van deze actie wordt geschat op basis van de data van elke dag. Dit gebeurt dus na 1 dag, na 2 dagen, na 3 dagen etc. door $\frac{N(t)}{t}$ te berekenen met *N(t)*: het respons aantal t/m dag t (bij keuze tijdseenheid 1 dag).

Op basis van de aantallen per dag, zijn schattingen $\hat{\lambda}(t)$ voor de parameter λ bepaald door $\frac{N(t)}{t}$ te berekenen.

9.5.1.2 Voorspellingsintervallen opstellen

Nu gaan we de totale respons voorspellen. Zoals in de theorie is besproken, is $24\,\hat{\lambda}(n)$ een schatting van de voorspelling van de totale respons. Dus op basis van de gevonden schattingen $\hat{\lambda}(n)$ kan $24\,\hat{\lambda}(n)$ bepaald worden voor elke *n*. Vervolgens kan dan met behulp van formule *(6)* het voorspellinginterval worden opgezet.. In grafiek 14 is het resultaat te zien.



Grafiek 15: *voorspelde totale respons met ongeveer 95% voorspellinginterval*

9.5.2 Model op actie A5727_B

Ook voor deze actie wordt de parameter λ geschat op basis van de data die elke dag binnenkomt. Daarna wordt de totale respons voorspeld na *n* dagen. Voor elke voorspelling wordt ook een voorspellingsinterval opgesteld. In de volgende paragrafen worden deze procedures besproken.

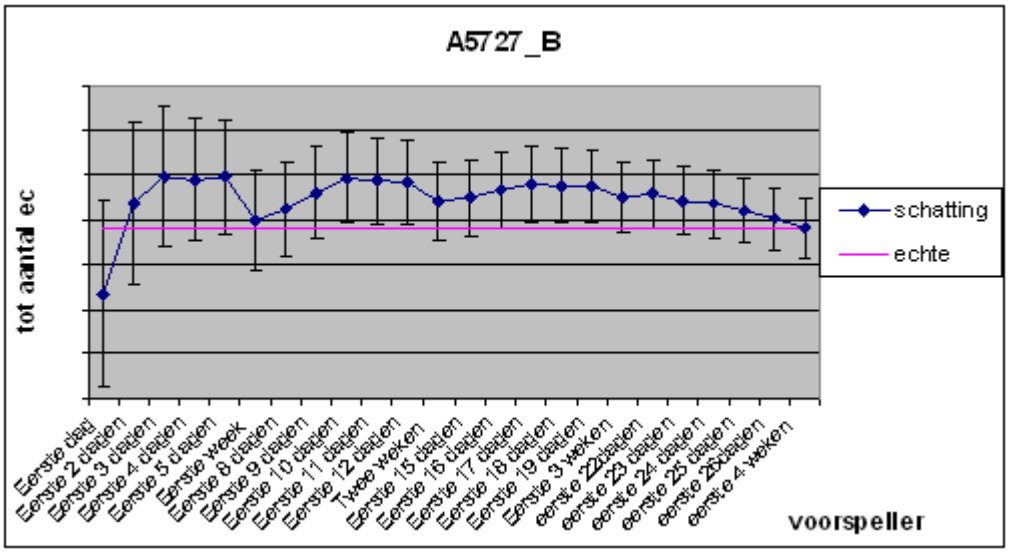
9.5.2.1 Schatten parameter λ

De parameter λ van deze actie wordt geschat op basis van de data van elke dag. Dit gebeurt dus na 1 dag, na 2 dagen, na 3 dagen etc. door $\frac{N(t)}{t}$ te berekenen met *N(t)* :het respons aantal t/m dag t (bij keuze tijdseenheid 1 dag).

Op basis van de aantallen per dag, zijn schattingen $\hat{\lambda}(t)$ voor de parameter λ bepaald door $\frac{N(t)}{t}$ te berekenen.

9.5.2.2 Intervallen

Ook voor deze actie wordt de totale respons voorspeld. Zoals in de theorie is besproken, is 24 $\hat{\lambda}(n)$ een schatting van de voorspelling van de totale respons. Dus op basis van de gevonden schattingen $\hat{\lambda}(n)$ uit tabel 24 kan 24 $\hat{\lambda}(n)$ bepaald worden voor elke *n*. Vervolgens wordt met behulp van de formules *(6)* het voorspellinginterval opgezet. In grafiek 15 is het resultaat te zien.



Grafiek15: *voorspelde totale respons met ongeveer 95% voorspellinginterval*

9.5.3 Model op actie A5727_C

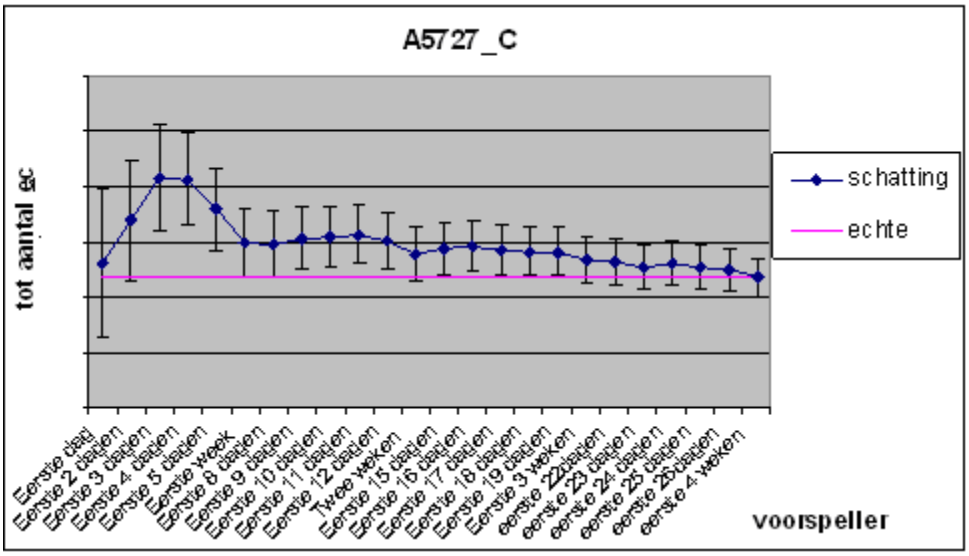
Als eerste wordt de parameter λ geschat op basis van de data die elke dag binnenkomt. Daarna wordt de totale respons voorspeld na *n* dagen. Voor elke voorspelling wordt ook het voorspellingsinterval opgesteld. In de volgende paragrafen worden deze procedures besproken.

9.5.3.1 Schatten parameter λ

De parameter λ van deze actie wordt geschat op basis van de data van elke dag.. Dit gebeurt dus na 1 dag , na 2 dagen, na 3 dagen etc. door $\frac{N(t)}{t}$ te berekenen
met *N(t)* :het respons aantal t/m dag t (bij keuze tijdseenheid 1 dag).
Op basis van de aantallen per dag, zijn schattingen $\hat{\lambda}(t)$ voor de parameter λ bepaald door $\frac{N(t)}{t}$ te berekenen.

9.5.3.2 Intervallen

Ook voor deze actie wordt de totale respons voorspeld. Zoals in de theorie is besproken, is 24 $\hat{\lambda}(n)$ een schatting van de voorspelling van de totale respons. Dus op basis van de gevonden schattingen $\hat{\lambda}(n)$ kan 24 $\hat{\lambda}(n)$ bepaald worden voor elke *n*. Vervolgens wordt met behulp van de formules *(6)* het voorspellinginterval opgezet. In grafiek 16 is het resultaat te zien.



Grafiek 16: *voorspelde totale respons met ongeveer 95% voorspellinginterval*

9.6 Conclusie

Uit de grafieken met de voorspellingsintervallen is te zien dat vanaf een bepaalde *n* de modellen het nog steeds niet erg goed doen. Er vindt namelijk altijd een overschatting plaats van de werkelijke totale respons. Wellicht is er dan toch enig trend in de data. Door het weinig aantal datapunten kan waarschijnlijk de witte ruis hypothese niet verworpen worden en is het aantal datapunten dus eigenlijk te klein om rigoureuze conclusies te kunnen trekken.

Een poisson proces is doorgaans een redelijk model voor een proces waarbij gebeurtenissen onafhankelijk van elkaar optreden in de tijd maar in dit geval is dit dus geen goede benadering voor de respons.

10 Onderzoek deel IV: Modelbouw op historische data

10.1 Inleiding

Gezien het resultaat van stationariteit uit het Tijdreeks onderzoek is een logisch vervolg om te onderzoeken hoe de respons op bepaalde dagen zich verhoudt tot de totale respons. Dus hoe de respons op de eerste dag, op de eerste twee dagen, op de eerste drie dagen etc. zich verhoudt tot de totale respons. Door historische data te gebruiken kunnen percentages van de totale respons per periode bepaald worden. Per periode wordt er dus bepaald hoeveel procentuele respons er voor de historische acties is binnengekomen. Met een periode wordt de eerste dag, eerste 2 dagen, eerst 3 dagen etc. bedoeld. Door deze model percentages op een nieuwe actie toe te passen kan bepaald worden welke periode de beste voorspeller voor de totale respons is.

De methodiek is om een model te bouwen op twee actie: op basis van deze twee actie worden modelpercentages bepaald per periode. Daarna wordt het model getest op een derde actie: Deze modelpercentages worden gebruikt om voor een nieuwe actie voorspellingen te geven voor het totaal aantal eerste contacten na elke periode. Omdat drie acties beschikbaar zijn worden er 3 (= $\binom{3}{2}$)
modellen gebouwd. Het bouwen van deze drie modellen op twee acties en het testen wordt in de paragrafen *Model 1, Model2* en *Model 3* beschreven.

Eerst worden een aantal begrippen verklaard die van toepassing zijn bij het evalueren van de gemaakte modellen. Deze twee definities zijn namelijk ondergrenzen die bij het evalueren van het model van toepassing kunnen zijn. Ook wordt beschreven hoe betrouwbaarheidsintervallen voor de schatting van de totale respons opgesteld kunnen worden.

10.2 Target

Elke actie heeft een target die vooraf wordt opgesteld. Deze target is voor het product hypotheken vaak een aantal ***netto*** hypotheken die de actie moet halen. Om dit netto aantal hypotheken terug te rekenen naar een aantal bruto hypotheken wordt gebruikt gemaakt van een *gemiddeld bruto-netto* percentage.

Gemiddelde Bruto-Netto percentage

In hoofdstuk 6.2.4 zijn de 95% betrouwbaarheidsintervallen opgesteld voor de bruto-netto percentages per actie. Het gemiddelde bruto-netto percentage wordt bepaald door het gemiddelde van de drie gemiddelde waardes van elk betrouwbaarheidsinterval te nemen. In tabel 26 zijn deze waardes te samengevat.

Het gemiddelde bruto-netto percentage is x %. Deze x % zal dus in de modellen gebruikt worden om de netto target terug te rekenen naar een bruto target. Na deze bruto target te hebben, moet deze nog terug gerekend worden naar een aantal bruto eerste contacten. Om deze berekening te maken, worden de conversie percentages van de Postbank gebruikt .

10.3 Ondergrens

Elke actie heeft een minimaal aantal klanten nodig om de kosten van het versturen van de mailings eruit te halen. Bij elke actie worden mailings verstuurd naar geselecteerde klanten. Aan deze mailing zitten kosten verbonden. Deze kosten zijn: *litho, druk, printen, ontwikkelingskosten, akt-materiaal, antwoordnummer, handlen , porto, adresaankoop* en *nabellen*.

Elke actie moet dus een minimaal aantal contacten (respondenten) behalen om te zorgen dat de kosten van de mailings eruit worden gehaald. Dit is dus een ondergrens (*break even point*) voor het te bouwen model.
Een kostenformule die deze *break even point* specificeert is:

Kosten=Opbrengst

Deze formule kan uitgeschreven worden naar:

*Omvang mailing * kosten per mailing = aantal respondenten * NCW[10] van 1 respondent*

Met het aantal respondenten wordt het aantal respondenten bedoeld die een hypotheek afsluiten. Deze kostenformule zal gebruikt worden om het minimum aantal respondenten te bepalen. Voor de NCW wordt een gemiddelde gebruikt over de drie acties.

Gemiddelde NCW
Elke actie heeft een gemiddelde NCW per respondent opgebracht. Door het gemiddelde van deze 3 NCW te nemen, ontstaat er dus een gemiddelde NCW. Deze wordt in de modellen gebruikt om het minimum aantal respondenten te bepalen.

De afgeronde gemiddelde NCW waarde zal in de kostenformule gebruikt worden voor het bepalen van het minimum aantal respondenten.

[10] NCW=Netto Contante Waarde.

10.4 Betrouwbaarheidsintervallen

Het is van belang voor elke schatting van de totale respons een betrouwbaarheidsinterval op te stellen. Op deze manier kan de kwaliteit van de voorspeller bepaald worden. Het echte totale aantal eerste contacten dat voorspeld moet worden moet namelijk, als een periode een goede voorspeller is, in het betrouwbaarheidsinterval opgenomen zijn.

Als eerst wordt de formule gegeven om een schatting van het totaal aantal eerste contacten te maken voor een nieuwe actie (actie 3) gebaseerd op de modelpercentages van twee historische acties (actie 1 en 2). De schatting voor de totale respons gebaseerd op de waargenomen respons in periode *i* van actie 3, wordt door de volgende formule gegeven:

$$TOT\_EC\_actie\ 3(periode\ i) = \frac{aantal\ contacten\ periode\ i\ \ actie\ 3}{mod elpercentage\ \ periode\ i} \qquad \textbf{(1)}$$

Met *TOT_EC__actie 3 (periode i)* wordt dus de schatting van het totaal aantal eerste contacten bedoeld gebaseerd op de respons aantallen in periode i van de nieuwe actie. Met periode *i* wordt de eerste *i* dagen bedoeld. Dus met periode 1 wordt dag 1 bedoeld, met periode 2 de eerste 2 dagen etc. Deze periode gaat tot maximaal n weken omdat je het aantal eerste contacten in de eerste n weken wil voorspellen.

Deze schatting bevat dus twee onzekerheden:

1. De onderzekerheid van de noemer van *(1):* het modelpercentage in periode i. Dit modelpercentage is een fractie die met behulp met de volgende formule is berekend op basis van twee historische acties, actie 1 en 2:

$$mod elpercentage\ \ periode\ i = \frac{aantal\ eerste\ contacten\ wa arg enomen\ in\ \ periode\ i\ van\ actie\ 1\ en\ 2}{totaal\ aantal\ wa arg enomen\ eerste\ contacten\ \ in\ eerste\ n\ weken\ actie\ 1\ en\ 2}$$

(2)

2. De onderzekerheid van de teller van (1). Dit aantal is slechts een waarneming van één steekproef (actie). Wat voor de nieuwe actie vanaf het begin bekend is, is de grootte van de mailgroep naar wie de actie is verstuurd. Het waargenomen aantal in periode *i* van actie 3 kan dan ook als een fractie worden gezien:

$$fractie\ actie\ 3\ \ periode\ i = \frac{aantal\ eerste\ contacten\ \ in\ \ periode\ i}{grootte\ \ mai lg roep\ \ actie\ 3}$$

In hoofdstuk 6.1.2 is al besproken dat de binomiale verdeling in nauwe relatie staat met het begrip fractie. Omdat voor grote *n* de normale verdeling gebruikt kan worden voor het benaderen van de binomiale verdeling, kan:

$$TOT\_EC\_actie\ 3(periode\ i) = \frac{aantal\ \ contacten\ \ periode\ \ i\ \ \ actie\ \ 3}{mod\ elpercentage\ \ \ periode\ \ i}$$

gezien worden als een deling van twee stochasten die beide normaal verdeeld zijn.

Omdat deze deling na enig literatuur onderzoek tot geen bestaande verdeling leidt, wordt het een complex probleem om een betrouwbaarheidsinterval voor deze schatting te bepalen. Een manier om deze complexiteit te omzeilen is door te simuleren. Zowel de modelpercentages als de aantallen van de nieuwe actie in periode i zijn gesimuleerd om zo een betrouwbaarheidsinterval op te stellen voor de schatting van het totaal gebaseerd op periode i. Op deze manier worden de twee onbetrouwbaarheden beide meegenomen bij het opstellen van het betrouwbaarheidsinterval.

Simuleren heeft alleen een nadeel. In praktijk hebben specialisten namelijk geen tijd voor simulaties. Een optie is dan door alleen de onbetrouwbaarheid van de waarneming van de nieuwe actie mee te nemen en zo een betrouwbaarheidsinterval voor het waargenomen aantal in periode i van een nieuwe actie op te stellen. Noem dit de *statistische methode*. De waarneming van de nieuwe actie kan, zoals al eerder beschreven, als een fractie worden geformuleerd:

$$fractie\ \ actie\ \ 3\ \ \ periode\ \ i = \frac{aantal\ \ eerste\ \ contacten\ \ \ \ in\ \ \ periode\ \ i}{grootte\ \ mai\,lg\ roep\ \ actie\ \ 3}$$

Rondom deze fractie kan een betrouwbaarheidsinterval worden opgesteld[11]. Dit leidt dan tot een betrouwbaarheidsinterval rondom het totaal aantal geschatte aantal eerste contacten. Dit wordt hieronder met een voorbeeld uitgelegd.

Voorbeeld
Stel dat op dag 1 van een nieuwe actie 27 klanten zijn waargenomen. Deze actie is naar 185.000 klanten verstuurd. De kans dat iemand op dag 1 respondeert is dan gelijk aan de volgende fractie

$$p: = \frac{27}{185000} = 0.0146\%.$$

Met behulp van de theorie uit hoofdstuk 6.1.2 kan voor deze fractie een betrouwbaarheidsinterval worden opgesteld. Deze is dan: [0.0091%, 0.0201%].

De *ondergrens* voor het aantal klanten op dag 1 is dan:

0.0091%* 185.000 = 17 klanten.

De *bovengrens* is dan 0.0091%* 185.000=38.
Een betrouwbaarheidsinterval voor het aantal klanten op dag 1 voor de nieuwe actie is dan [17 38].

Stel dat het modelpercentage op dag 1 3,84% is. Dat wil zeggen dat 3,84% van de totale respons op dag 1 is binnengekomen Dit percentage is bepaald door voor twee historische acties te kijken hoeveel procent er op dag 1 binnenkwam.

[11] In hoofdstuk 6.1.2 is de theorie over het opstellen van een betrouwbaarheidsinterval rondom een fractie te vinden.

Met behulp van formule (1) kan nu een schatting voor het totaal aantal eerste contacten bepaald worden:

Gebaseerd op de 27 klanten op dag 1 is een schatting voor het totaal aantal eerste contacten:
$\frac{27}{3{,}84\%} = 703$ eerste contacten.
De ondergrens voor het totaal aantal eerste contacten kan dan bepaald worden door de ondergrens van 16,8 klanten op dag 1 te nemen: : $\frac{17}{3{,}84\%} = 443$

De bovengrens voor het totaal aantal eerste contacten kan dan bepaald worden door de bovengrens van 37,2 klanten op dag 1 te nemen: $\frac{38}{3{,}84\%} = 989$

Een betrouwbaarheidsinterval voor het totaal aantal eerste contacten , gebaseerd op de eerste dag van de actie is dan: [443 , 989]

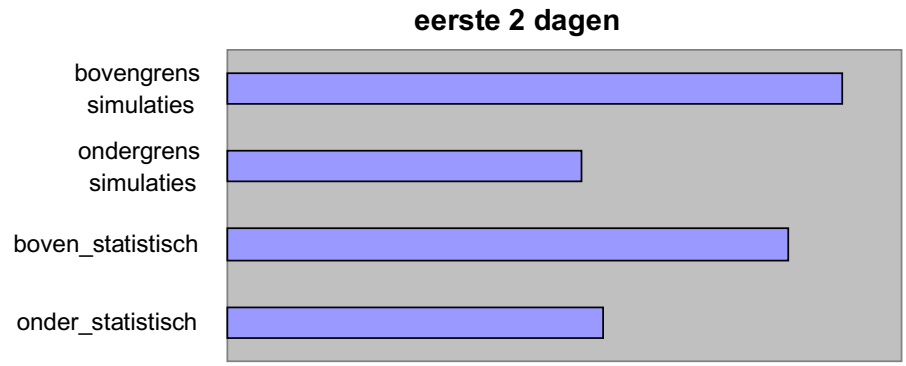
Op deze statistische manier kan voor alle periodes een betrouwbaarheidsinterval worden opgesteld voor het totaal aantal eerste contacten.

Verschil twee soorten betrouwbaarheidsintervallen

Met behulp van *simulaties* is een bovengrens en ondergrens bepaald voor het totale aantal eerste contacten geschat met behulp van de data van de eerste twee dagen. De code van deze simulaties is in bijlage G te vinden. In dit verkregen interval zitten de twee onzekerheden bevat. De onderzekerheid van het modelpercentage en de onzekerheid van het aantal eerste contacten op de eerste twee dagen van de nieuwe actie. Ook is een bovengrens en ondergrens bepaald door alleen de onzekerheid van het aantal eerste contacten op de eerste twee dagen van de nieuwe actie mee te nemen, de *statistische methode*.

In grafiek 20 zijn de ondergrenzen en bovengrenzen van de betrouwbaarheidsintervallen voor de schatting van het totaal aantal eerste contacten gebaseerd op eerste 2 dagen van de nieuwe actie te zien. Het ene interval is door simulaties bepaald en het andere door de statistische methode toe te passen.

Wat uit deze grafiek valt te zien is dat het simulatie interval iets breder is dan het interval dat op de statistische manier bepaald is. Dit verschil in breedte heeft te maken dat bij het simuleren ook de onzekerheid van het modelpercentage is meegenomen waardoor je dus een breder interval krijgt. Het kleine verschil in breedte komt dus door onbetrouwbaarheid die door het modelpercentage wordt veroorzaakt.



Grafiek 20: *Verschil in betrouwbaarheidsintervallen simuleren en statistische methode*

Conclusie

Voor het opstellen van betrouwbaarheidintervallen voor het geschatte totaal aantal eerste contacten wordt in dit onderzoek niet gebruik gemaakt van simulaties maar van de statistische manier. Hier wordt een betrouwbaarheid rondom het aantal in periode i van de nieuwe actie opgesteld . Dit kan door dit aantal als fractie te zien.

Het verschil in betrouwbaarheidsintervallen is relatief klein. Het verkregen interval is op deze manier iets kleiner waardoor eerder een aanname wordt gedaan dat periode i geen goede schatter voor het totaal aantal eerste contacten omdat de echte waarde niet binnen dat interval valt. Je krijgt op deze manier dus een strakkere grens wat alleen kan leiden tot het eerder concluderen dat een periode geen goede voorspeller is voor het totaal.

10.5 Model 1

Dit model is gebouwd op de acties A5727_B en A5727_C. Deze twee acties zijn samengevoegd en vervolgens is er bepaald hoeveel respons er totaal op de eerste dag (som van wat er op dag1 bij actie A5727_B en actie A5727_C binnenkomt), eerste twee dagen etc. binnenkomt.

10.5.1 Modelbouw

Voor elke periode is een responspercentage bepaald. Dus hoeveel respons er de eerste dag binnenkomt ten opzichte van het totale aantal waargenomen eerste contacten.

10.5.2 Testen model

Met behulp van de modelpercentages kan nu voor een nieuwe actie de totale respons voorspeld worden. Dit model wordt toegepast op actie A6144.

Schatten totale respons

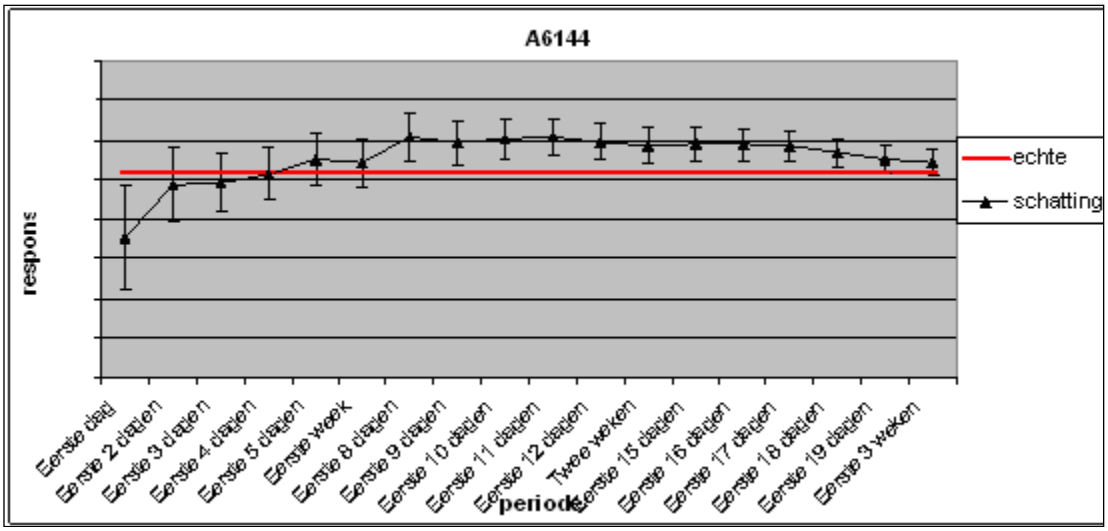
Op basis van de responspercentages van dit model kan nu een schatting worden gegeven van het totaal aantal eerste contacten na een maand. Dit wordt voor de eerste periode, de eerste dag, uitgelegd.

Voorbeeld
Vertrouwelijk

Dit wordt voor alle periodes gedaan. Het resultaat is te zien in grafiek 21 waar het echte aantal eerste contacten en het geschatte aantal eerste contacten per periode te zien zijn.

Betrouwbaarheidsinterval respons

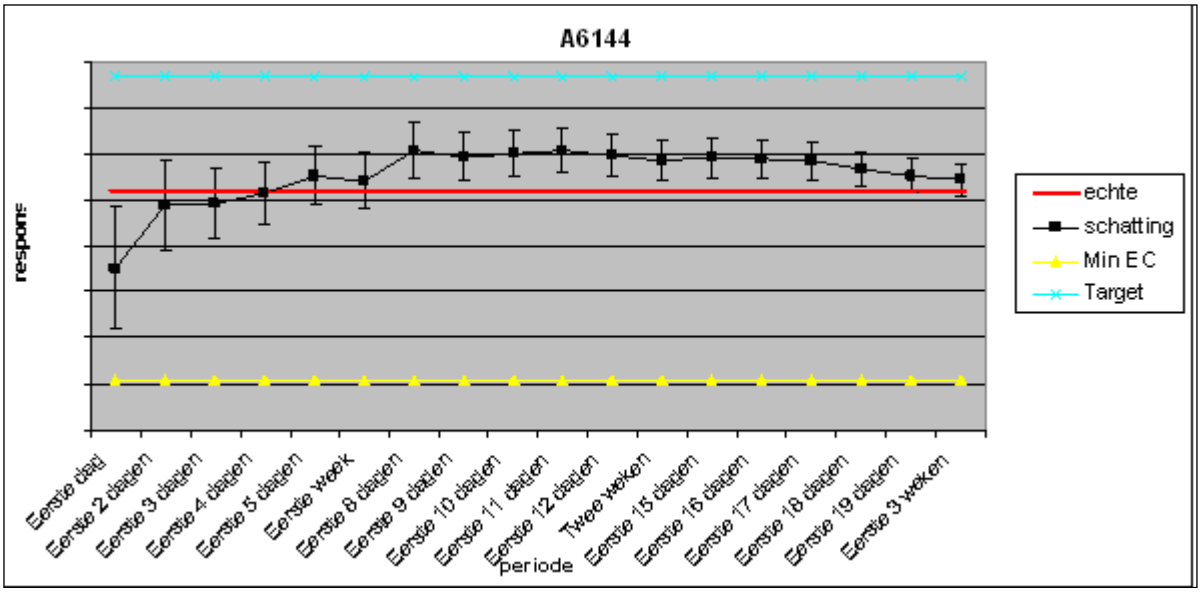
Voor elk geschat aantal eerste contacten kan een betrouwbaarheidsinterval worden opgesteld. Dit door de statistische methode te gebruiken. In het voorbeeld op blz.65 is een betrouwbaarheidsinterval opgesteld voor het totaal aantal eerste contacten gebaseerd op de respons in periode 1 (dag 1) van de nieuwe actie. Door deze procedure voor elke periode i te herhalen kan voor elk geschat aantal eerste contacten een betrouwbaarheidsinterval worden opgesteld. Het resultaat hiervan is in grafiek 21 te zien.



Grafiek 21: *Schattingen totaal aantal eerste contacten actie A6144*

Grafiek

Na de target en het minimaal aantal eerste contacten (*min EC*) berekend te hebben, kunnen deze waardes in een grafiek worden weergeven samen met de geschatte en waargenomen respons. In grafiek 22 is het resultaat te zien.



Grafiek 22: *Schattingen actie A6144 met ondergrens en target*

10.5.3 Conclusie

Dit model geeft weer dat vanaf de tweede dag al een goede schatting van de totale respons gegeven kan worden. Na een week wordt de schatting echter onnauwkeurig.

10.6 Model 2

Dit model is gebouwd op de acties A5727_A en A5727_B. Deze twee acties zijn samengevoegd en dan is er bepaald hoeveel respons er totaal op de eerste dag (som van wat er op dag 1 bij actie A5727_B en actie A5727_A binnenkomt), eerste twee dagen etc. binnenkomt.

10.6.1 Modelbouw

Als eerste wordt voor elke periode de modelpercentages bepaald. Dus hoeveel respons er de eerste dag binnenkomt ten opzichte van het totale aantal eerste contacten.

10.6.2 Testen Model

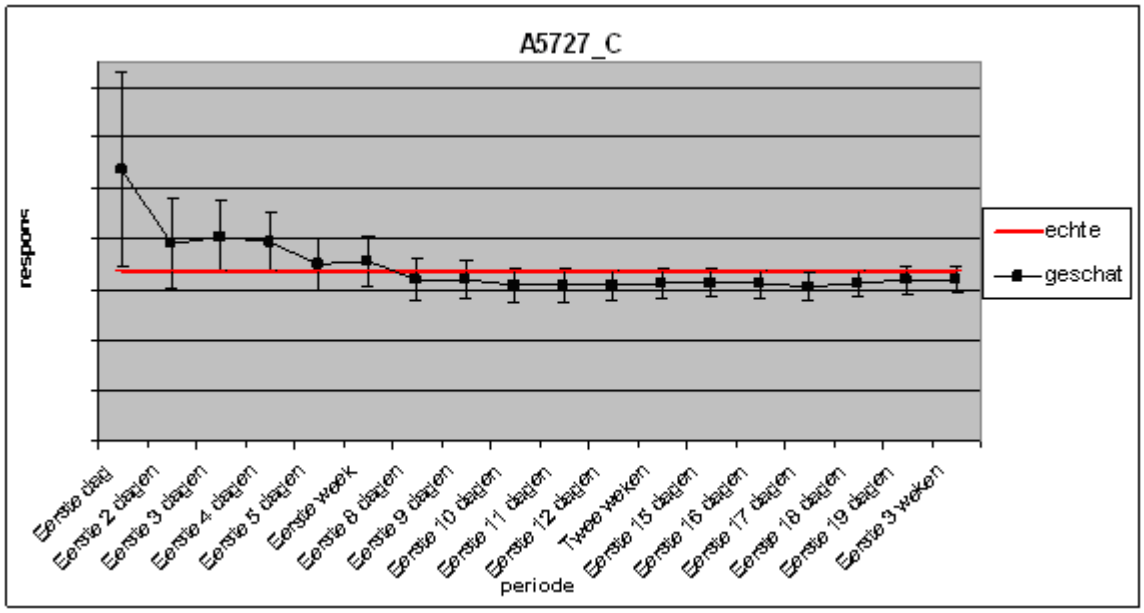
Met behulp van de modelpercentages kan nu voor een nieuwe actie de totale respons voorspeld worden. Dit model wordt toegepast op actie A5727_C.

Schatten totale respons
Op basis van de responspercentages van dit model kan dus een schatting worden gegeven van het totaal aantal eerste contacten na een maand. Dit wordt voor de eerste periode, de eerste dag, uitgelegd.

Dit wordt voor alle periodes gedaan. Het resultaat is te zien in grafiek 23 waar het echte aantal eerste contacten en het geschatte aantal eerste contacten per periode te zien zijn.

Betrouwbaarheidsinterval geschatte respons

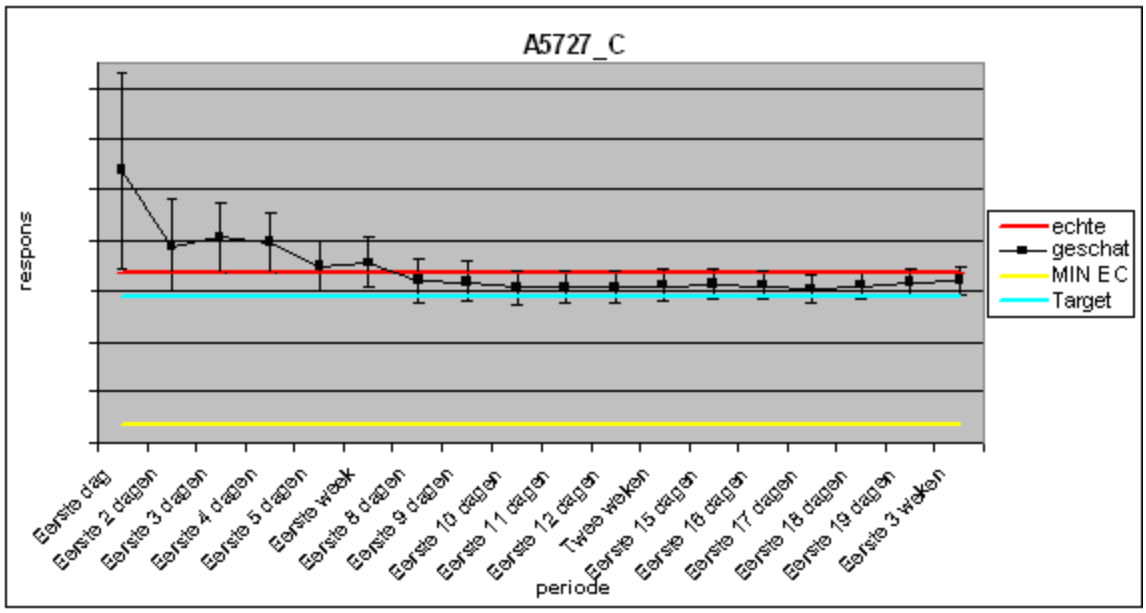
Voor elk geschat aantal eerste contacten kan een betrouwbaarheidsinterval worden opgesteld. Dit door de statistische methode te gebruiken. In het voorbeeld op blz. 65 is een betrouwbaarheidsinterval opgesteld voor het totaal aantal eerste contacten gebaseerd op de waargenomen respons in periode 1 (dag 1) van de nieuwe actie. Door deze procedure voor elke periode i te herhalen kan voor elk geschat aantal eerste contacten een betrouwbaarheidsinterval worden opgesteld. Het resultaat hiervan is in grafiek 23 te zien.



Grafiek 23: *Schattingen totaal aantal eerste contacten actie A5727_C*

Grafiek

Na de target en het minimaal aantal eerste contacten (min EC) berekend te hebben, kunnen deze waardes in een grafiek worden weergeven samen met de geschatte en waargenomen respons. In grafiek 24 is het resultaat te zien.



Grafiek 24: *Schattingen actie A5727_C met ondergrens en target*

10.6.3 Conclusie

Dit model geeft weer dat in periode 2, de tweede dag na de mailing, al een goede schatting van de totale respons gegeven kan worden. De derde en vierde dag bevatten net aan de echte waarde maar daarna wordt altijd een goede schatting gegeven met een niet al te groot betrouwbaarheidsinterval. Over het algemeen geeft dit model altijd een goede schatting van de totale respons.

10.7 Model 3

Dit model is gebouwd op de acties A5727_A en A5727_C. Deze twee acties zijn samengevoegd en dan is er bepaald hoeveel respons er totaal binnenkomt op de eerste dag (som van wat er op dag1 bij actie A5727_C en actie A5727_A binnenkomt), eerste twee dagen etc.

10.7.1 Modelbouw

Als eerste wordt voor elke periode de modelpercentages bepaald. Dus hoeveel respons er de eerste dag binnenkomt ten opzichte van het totale aantal eerste contacten.

10.7.2 Testen Model

Met behulp van de modelpercentages kan nu voor een nieuwe actie de totale respons voorspeld worden. Dit model wordt toegepast op actie A5727_B

Schatten totale respons

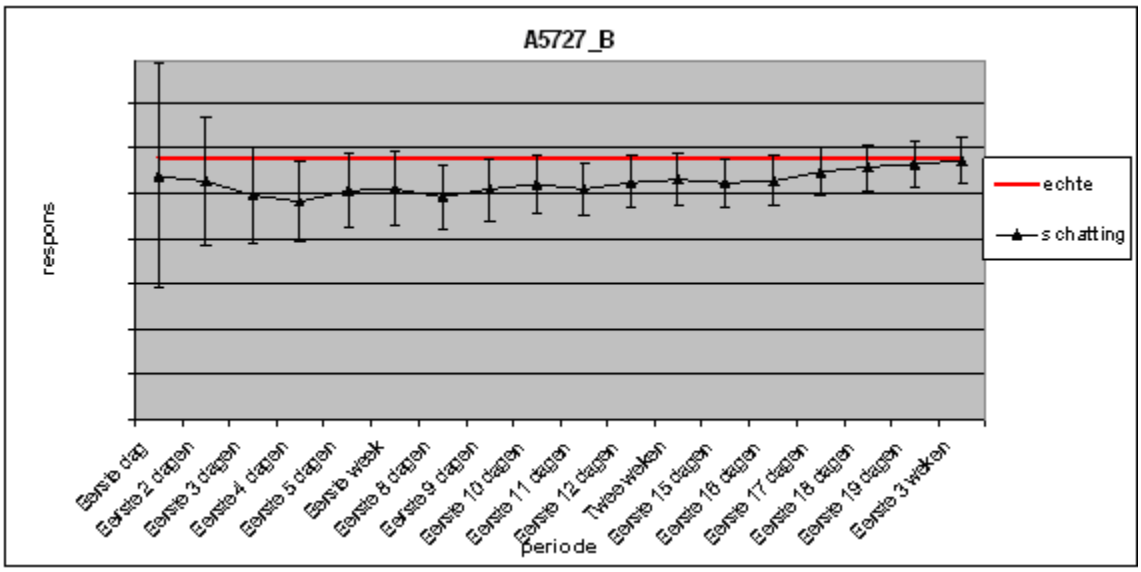
Op basis van de responspercentages van dit model kan dus een schatting worden gegeven van het totaal aantal eerste contacten na een maand. Dit wordt voor de eerste periode, de eerste dag, uitgelegd.

Voorbeeld
Vertrouwelijk

Dit wordt voor alle periodes gedaan. Het resultaat is te zien in grafiek 25 waar het echte aantal eerste contacten en het geschatte aantal eerste contacten per periode te zien zijn.

Betrouwbaarheidsinterval geschatte respons

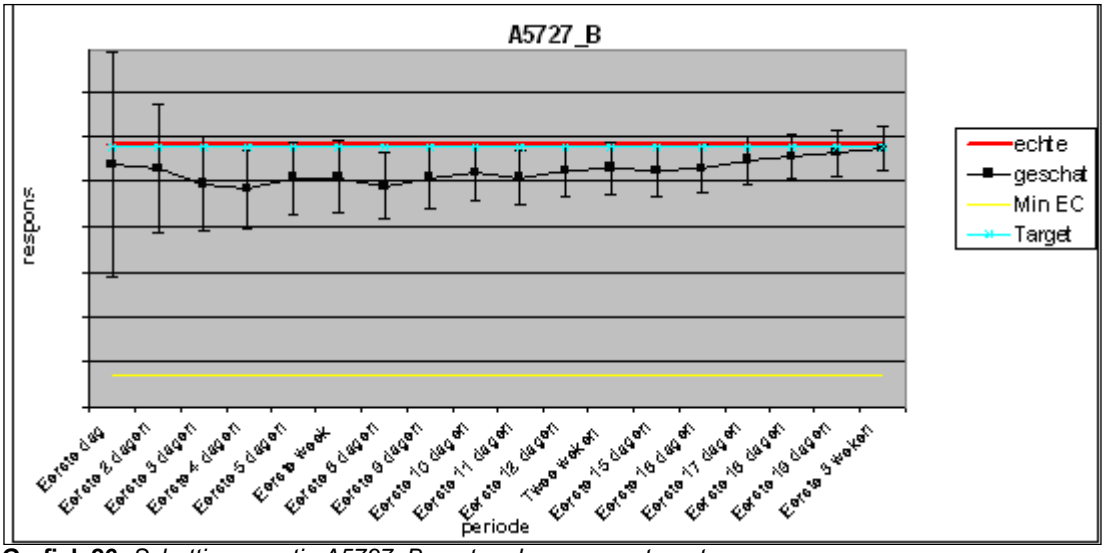
Voor elk geschat aantal eerste contacten kan een betrouwbaarheidsinterval worden opgesteld. Dit door de statistische methode te gebruiken. In het voorbeeld op blz. 65 is een betrouwbaarheidsinterval opgesteld voor het totaal aantal eerste contacten gebaseerd op de waargenomen respons in periode 1 (dag 1) van de nieuwe actie. Door deze procedure voor elke periode i te herhalen kan voor elk geschat aantal eerste contacten een betrouwbaarheidsinterval worden opgesteld. Het resultaat hiervan is in grafiek 25 te zien.



Grafiek 25: *Schattingen totaal aantal eerste contacten actie A5727_B*

Grafiek

Na de target en het minimaal aantal eerste contacten (*min EC*) berekend te hebben, kunnen deze waardes in een grafiek worden weergeven samen met de geschatte en waargenomen respons. In grafiek 26 is het resultaat te zien.



Grafiek 26: *Schattingen actie A5727_B met ondergrens en target*

10.7.3 Conclusie

Uit de grafiek valt te zien dat dit model altijd een onderschatting geeft van de echte waarde. De betrouwbaarheidsintervallen bevatten wel de echte waarde van de totale respons. Ook hier geldt dat vanaf de tweede dag al een goede schatting voor de totale respons kan worden gegeven.

10.8 Eind Conclusie

Op basis van de drie modellen kan vanaf *n* dagen al een goede voorspelling voor de totale bruto respons gemaakt worden. Op basis van deze geschatte aantallen kan dan door middel van de conversie percentage van de Postbank bepaald worden hoeveel de bruto productie is die hieruit volgt. Om vervolgens de netto productie te bepalen, kan gebruikt worden gemaakt van het *bruto-netto* percentage.

Deze netto productie is dan de netto productie die volgt uit de contacten van de eerste maand. De gevonden methode wordt stapsgewijs in bijlage H beschreven.

11 Resultaten

Hieronder zijn per onderzoeksonderdeel de resultaten kort en bondig opgeschreven.

Vooronderzoek productie actie
Dit vooronderzoek is uitgevoerd om te onderzoeken waar de productie van de actie voor het product hypotheken zit. De onderzoeksvraag is als volgt geformuleerd:

In welke periode zit de productie van de actie?

Het vinden van deze productie is op twee verschillende manieren onderzocht: met de *Waterval* methode en de *Directe* methode. Na het toepassen van deze twee methodes volgt dat de eerste *n* weken van de eerste contacten een goede benadering geven van de hypotheekproductie die getriggered is door de actie. Deze productie bevat niet de tussenpersonen productie.

Modelbouw op gemiddeldes van historische data
De onderzoeksvraag is bij dit deel als volgt geformuleerd:

Welke periode (dus de eerste dag, de eerste twee dagen etc.) is de beste voorspeller voor de totale respons?

Op basis van de drie verkregen modellen kan al n dagen na de mailing, een goede schatting worden gegeven van het totaal aantal bruto eerste contacten. Op basis van deze geschatte aantallen kan dan door middel van conversie percentages worden bepaald hoeveel de bruto productie is die hieruit volgt. Om vervolgens de netto productie te bepalen kan gebruik worden gemaakt van het *bruto-netto* percentage, deze is de productie van het Postbank kanaal.

De methode is in de bijlage H terug te vinden.

Tijdreeksen
Voor dit deel is de onderzoeksvraag als volgt geformuleerd:

Kan het responsverloop van de eerste contacten door middel van tijdreeks modellen gemodelleerd worden?

Het responsverloop van de eerste n weken van de eerste contacten kan, na het uitvoeren van stationariteit en witte ruis toetsen, worden gezien als een stationair witte ruis proces. Dit volgt voor twee acties . Een witte ruis proces is een reeks van ongecorreleerde random variabelen, waarvan de verwachting en de variantie constant is. Met andere woorden, de waargenomen respons aantallen per dag zijn identiek verdeelde onafhankelijke random variabelen en kunnen dus niet door een tijdreeks model gemodelleerd worden. Overigens is het aantal datapunten waarop deze witte ruis conclusie is gebaseerd klein.

Classificatie Klanten

De onderzoekvraag is voor dit onderdeel als volgt:

Zijn er klanten met bepaalde eigenschappen die altijd vroeg of laat responderen? M.a.w Kunnen klanten geclassificeerd worden op basis van hun responstijd?

Na meervoudige lineaire regressie toegepast te hebben, duiden de gevonden slechte determinatiecoëfficiënten en liftcurves erop dat er geen meervoudige verbanden zijn. Er kan dus geconcludeerd worden dat de variabele *responstijd* ***geen verband*** heeft met de beschikbare variabele. De *responstijd* van een klant is dus een variabele die niet voorspeld kan worden op basis van klant kenmerken.

Modelbouw door Poisson benadering

De modellen verkregen door het responsverloop als een Poisson proces te modelleren doen het na enige *n* nog steeds niet heel goed. Dit zou kunnen betekenen dat er toch enige trend is in de data. Waarschijnlijk kan wegens het kleine aantal datapunten de witte ruis hypothese niet worden verworpen en kunnen hieruit dus geen rigoureuze conclusies getrokken worden.

12 Conclusie & Aanbevelingen

12.1 Conclusie

De voornaamste conclusies die op basis van de vier onderzoeken gemaakt kunnen worden, zijn hieronder beschreven.

- **De productie uit maand n van de eerste contacten geeft een goede benadering voor de totale productie.**
  Dit volgt na het toepassen van twee verschillende methodes om de productie van de actie te bepalen.

- **Een model gebouwd op gemiddeldes van historische data kan al na 2 dagen goede voorspellingen geven voor de totale respons.**
  Het bouwen van modellen op historische data, geeft goede benaderingen voor de totale verwachte respons van een nieuwe actie. Dit door gemiddelde responspercentages te nemen van binnengekomen respons per dag van historische acties. De methodiek is om een model te bouwen op twee historische acties en deze te testen op een nieuwe derde actie. Door de verkregen modellen ontstaat er een beeld van welke dagen de beste voorspellers zijn voor het totaal aantal eerste contacten in de eerste maand.

  Uit dit onderzoek volgt dus dat vanaf twee dagen na de actie, al een goede schatting kan worden gemaakt van het totaal aantal bruto eerste contacten. Op basis van deze geschatte aantallen kan dan door middel van conversie percentages bepaald worden hoeveel de bruto productie is die hieruit volgt. Om vervolgens de netto productie te bepalen, kan gebruik worden gemaakt van *bruto-netto* percentages

- **Het responsverloop van de eerste n weken van de eerste contacten van het product hypotheken is een stationair witte ruis proces.**
  Dit houdt in dat de waargenomen respons aantallen per dag identiek verdeelde en onafhankelijke random variabelen zijn en dus niet door tijdreeks modellen gemodelleerd kunnen worden. Overigens is het aantal datapunten waarop deze witte ruis conclusie is gebaseerd klein (alleen de eerste n weken).

- **Responstijd is een variabele die niet voorspeld kan worden op basis van de gebruikte variabelen.**
  Na het uitvoeren van lineaire regressie, tonen de resultaten uitzonderlijk slechte liftcurves en determinatiecoëfficiënten. Je kunt op basis hiervan concluderen dat de variabele *responstijd* ***geen verband*** heeft met de beschikbare variabele. De *responstijd* van een klant is dus een variabele die niet voorspeld kan worden op basis van klant kenmerken.

- **Modelleren van de binnenkomende respons als een Poisson proces is geen goed model.**
  Dit komt waarschijnlijk doordat er toch enige trend in de data is. Het aantal datapunten waarop de aanname van witte ruis is aangenomen is waarschijnlijk te klein om rigoureuze conclusies te kunnen trekken.

12.2 Aanbevelingen

Hieronder worden enkele aanbevelingen voor de gevonden methode en mogelijke vervolgonderzoeken gegeven.

❖ Gevonden methode

Het model kan naarmate er meer data voor handen is bijgewerkt en ge-update worden. Bepaalde aannames of opgestelde betrouwbaarheidsintervallen die in dit onderzoek gemaakt zijn, moeten naarmate er meer data beschikbaar is, weer getoetst worden. Dit heeft betrekking op de volgende kwesties:

- **Alleen in maand n van de eerste contacten zit het effect van de actie.**

Naarmate er meer data is, kan onderzocht worden of deze n weken wellicht minder of meer kunnen worden.

- **Betrouwbaarheidsintervallen om bruto-netto percentages door middel van bootstrapping.**

Door middel van bootstrappen zijn er betrouwbaarheidsintervallen om de bruto-netto percentages bepaald. Naarmate er meer data is, kunnen deze betrouwbaarheidsintervallen veranderen waardoor deze bijgewerkt moeten worden.

❖ Vervolgonderzoek

Hieronder enkele aanbevelingen voor mogelijke vervolgonderzoeken.

- **Andere Postbank producten**

Helaas was er geen tijd meer beschikbaar om andere producten van de Postbank te onderzoeken. Enkele interessante vraagstukken voor andere producten zijn:

1. In welke periode zit de productie van een actie voor andere producten?
2. Is het responsverloop ook voor andere producten witte ruis?

- **Responstijd**
  In dit onderzoek is geconcludeerd dat de gebruikte variabelen, geen invloed hebben op de variabele *responstijd*. Interessant zou zijn om te onderzoeken welke factoren *wel* invloed op deze variabele hebben.

- **Responsperiode Postbank**
  Vertrouwelijk

- **Modelleren respons proces door niet-homogeen Poisson proces**
  Dit is een andere aanpak die wegens gebrek aan tijd niet is uitgevoerd. In dit geval is de Poisson parameter geen constante maar kan in de tijd veranderen. Het verwacht aantal gebeurtenissen tussen *a* en *b* wordt dan weergegeven door de functie:

$$\lambda_{a,b} = \int_a^b \lambda(t)\, dt.$$

- **Tool bouwen**
  Tenslotte is een logisch vervolg op dit onderzoek, het bouwen van een tool die de gevonden methode implementeert.

13 Referenties

Buijs, A. (1999)*; Statistiek om mee te werken*; Educatieve Partners Nederland

Buijs, A. (2000); *Statistiek om mee verder te werken*; Stenfert Kroese

Chatfield, C. (2004); *The Analysis of Time Series*; Chapman & Hall/CRC

Karg, Y. (2004); *Beslisboom of Regressie?* ; Afstudeerscriptie Bedrijfswiskunde & Informatica, Vrije Universiteit, Amsterdam.

Koole, G. (2007); O*ptimization of Business Processes: An Introduction to Applied Stochastic Modelling*; Collegedictaat , Vrije Universiteit

Grimmett, G.R, (1992); *Probability and Random Processes*; Oxford Science Publications

Gunst, M. de, (2006)*; Statistical Models*; Collegedictaat; Vrije Universiteit, Amsterdam

Gunst, M. de, Vaart van der A.W. (2005); *Statistische Data Analyse*; Collegedictaat Vrije Universiteit, Amsterdam;

Pindyck R, Rubinfield D. (1999); *Econometric Models & Economic Forecasts;* McGraw-Hill INC.

Rid, O.P (2001); *Data Mining Cookbook*; Wiley Computer Publishing

Tip, W. e.a. *; Slim testen met Vet rendement;* uitgave Postbank

Vaart, A. van der (2003); *Algemene Statistiek*; Collegedictaat; Vrije Universiteit

Witten, I., Frank E. (2005); *Data Mining; Practical Machine Learning Tools and Techniques*; Morgan Kaufmann Publishers

ING

Bijlage C: Onderzoek deel I

Output R: stationariteit & witte ruis toetsen en autocorrelograms

Stationariteit toetsen per actie

Actie A5727_C

Met zaterdagen

```
        Phillips-Perron Unit Root Test

data:  c
Dickey-Fuller = -3.9233, Truncation lag parameter = 2, p-value = 0.02690
```

Zonder Zaterdagen

```
        Phillips-Perron Unit Root Test

data:  c_zz
Dickey-Fuller = -5.2334, Truncation lag parameter = 2, p-value = 0.01
```

Actie A5727_B

Met zaterdagen

```
        Phillips-Perron Unit Root Test

data:  b
Dickey-Fuller = -4.1597, Truncation lag parameter = 2, p-value = 0.01768
```

Zonder Zaterdagen

```
        Phillips-Perron Unit Root Test

data:  b_zz
Dickey-Fuller = -3.0475, Truncation lag parameter = 2, p-value = 0.1733
```

Actie A5727_A

Met Zaterdagen

```
        Phillips-Perron Unit Root Test

data:  a
Dickey-Fuller = -5.1384, Truncation lag parameter = 2, p-value = 0.01
```

Zonder Zaterdagen

```
        Phillips-Perron Unit Root Test

data:  a_zz
Dickey-Fuller = -4.4985, Truncation lag parameter = 2, p-value = 0.01
```

ING

Witte ruis toetsen per actie

Actie A5727_B

Differenced data van actie A5727_B

```
  Phillips-Perron Unit Root Test

data:  diff(b_zz)
Dickey-Fuller = -7.0108, Truncation lag parameter = 2, p-value = 0.01
```

Actie A6144

```
Box.test(a, type="Ljung-Box")

        Box-Ljung test

data:  a
X-squared = 0.9098, df = 1, p-value = 0.3402
```

```
Box.test(a_zz, type="Ljung-Box")

        Box-Ljung test

data:  a_zz
X-squared = 5.4712, df = 1, p-value = 0.01933
```

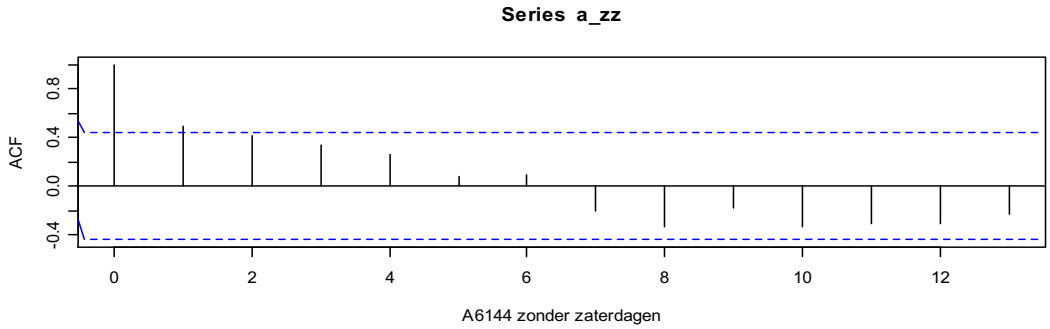
Actie A5727_C

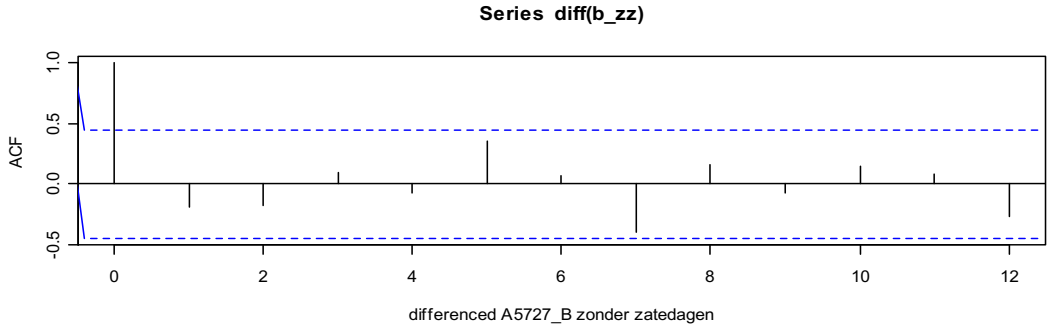
```
Box.test(c_zz, type="Ljung-Box")

        Box-Ljung test

data:  c_zz
X-squared = 0.757, df = 1, p-value = 0.3843
```

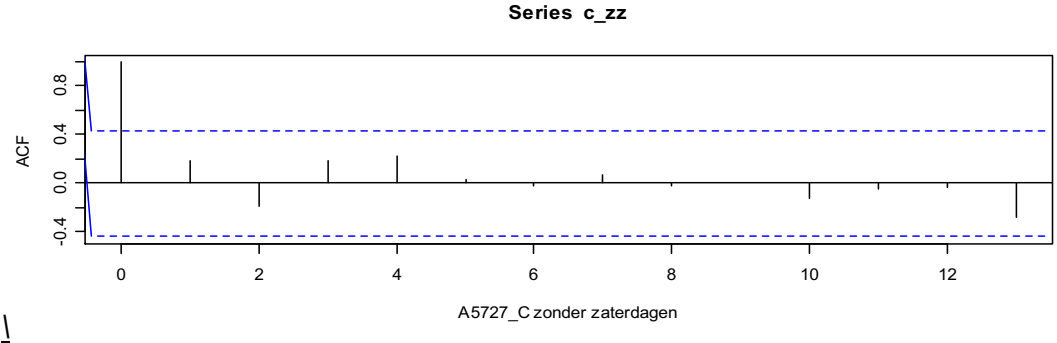
```
Box.test(c, type="Ljung-Box")
        Box-Ljung test
data:  c
X-squared = 1.4697, df = 1, p-value = 0.2254
```

ING

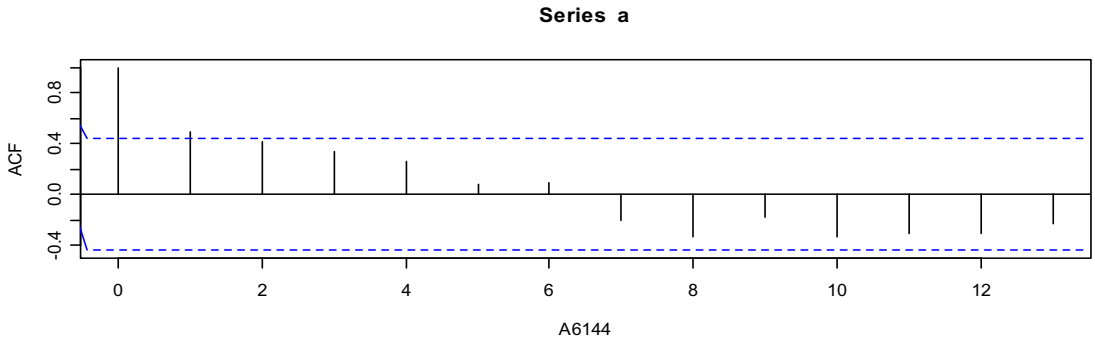
Autocorrelograms

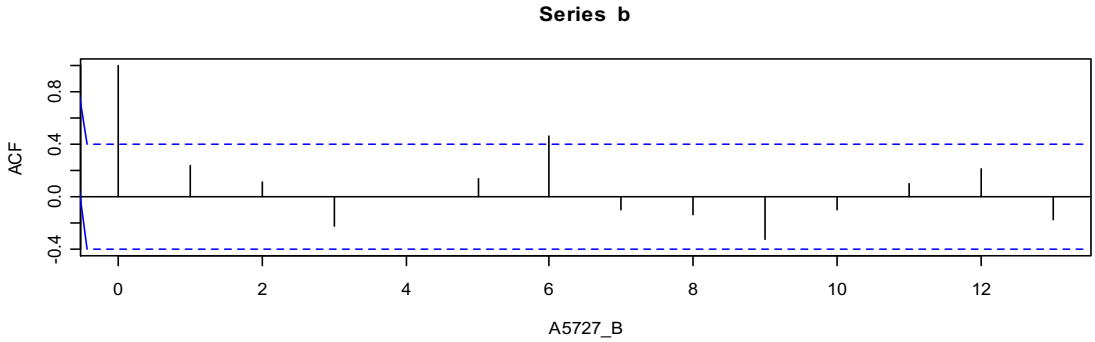


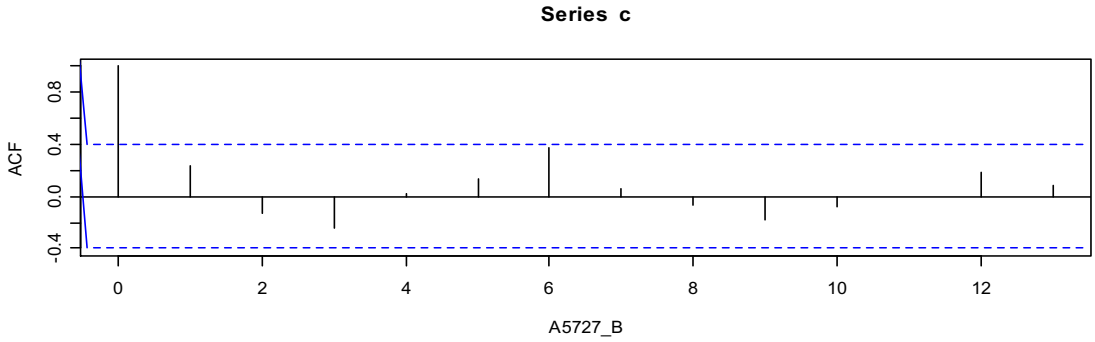




\


Series a
ACF
0.8
0.4
0.0
-0.4
0
2
4
6
8
10
12
A6144

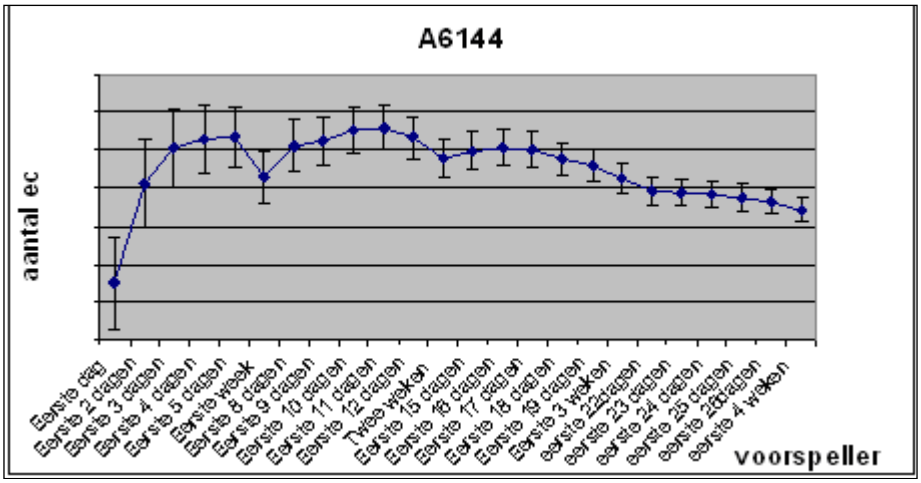

Series b
ACF
0.8
0.4
0.0
-0.4
0
2
4
6
8
10
12
A5727_B


Series c
ACF
0.8
0.4
0.0
-0.4
0
2
4
6
8
10
12
A5727_B

Bijlage D: Onderzoek deel III

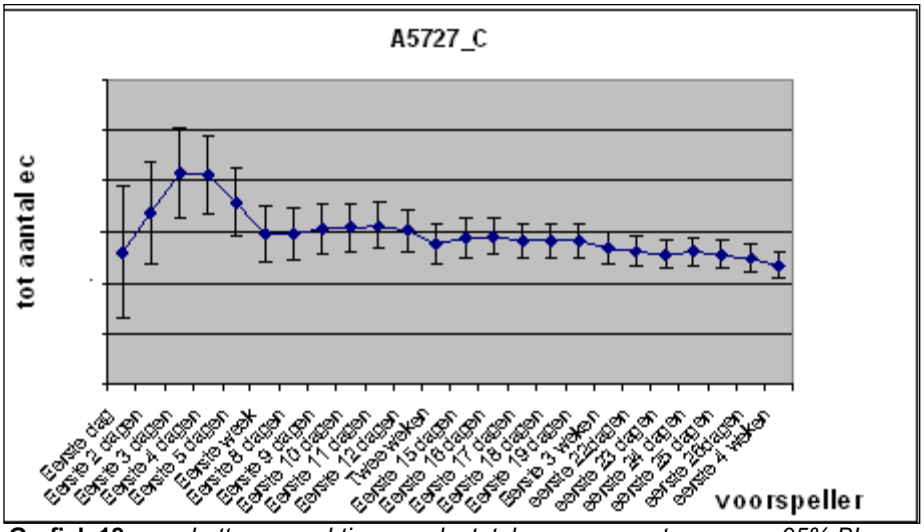
schatting poisson parameter, betrouwbaarheidsintervallen en voorspellingsintervallen

Actie A6144



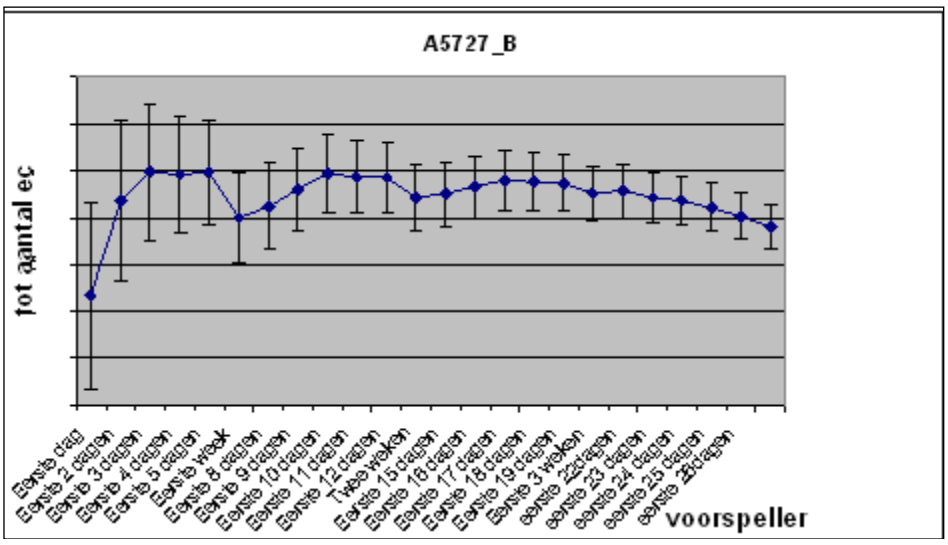
Grafiek 4b: *Geschatte verwachting van de totale respons met 95% BI*

Actie A5727_C



Grafiek 18: *geschatte verwachting van de totale respons met ongeveer 95% BI*

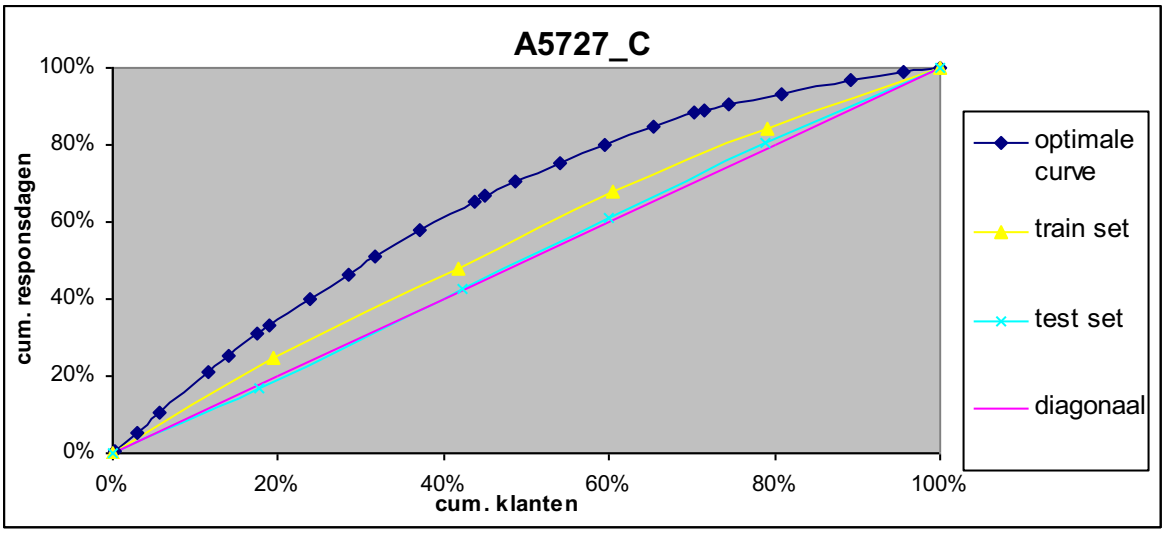
Actie A5727_B



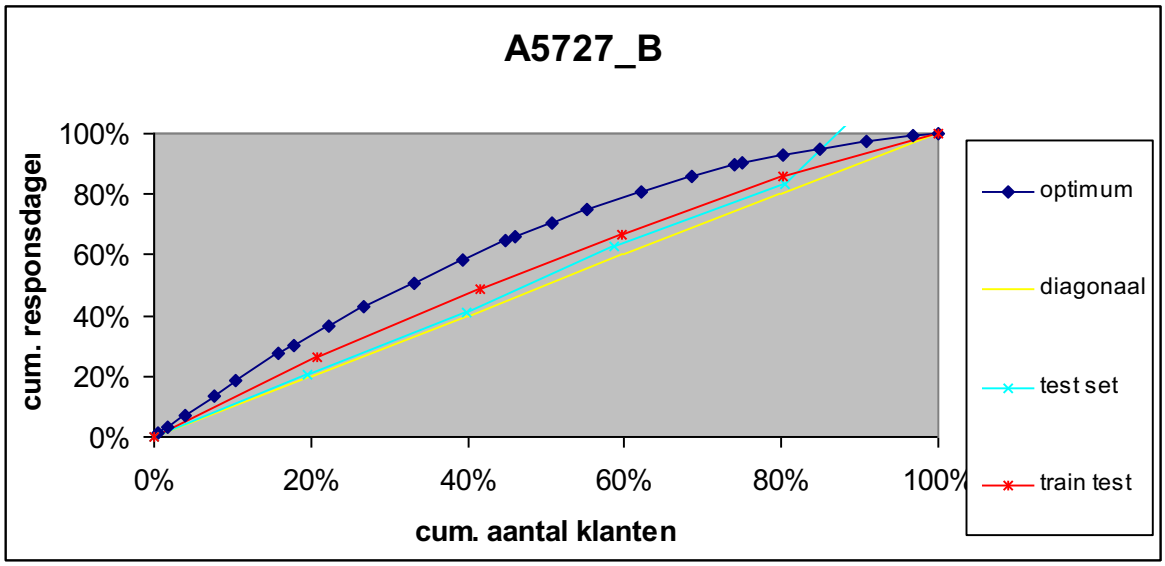
Grafiek 16: *geschatte verwachting van de totale respons met ongeveer 95% BI*

Bijlage E: Onderzoek deel II

Liftcurve Actie A5727_C



Liftcurve Actie A5727_B



Bijlage F: Onderzoek deel II

Output Stepwise Regression

A6144

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars	Partial R-Square	Model R-Square	F Value	Pr > F
1		1	0.0174	12.69	0.0004		
2		2		0.0290	8.51		0.0036
3		3		0.0404	8.55		0.0036
4		4		0.0492	6.53		0.0108
5		5		0.0569	5.75		0.0167
6		6		0.0636	5.11		0.0241

A5727_B

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1		1	0.0338	0.0338	57.6945	14.33	0.0002	
2		2	0.0304	0.0642	45.0230	13.30	0.0003	
3		3	0.0216	0.0858	36.6041	9.65	0.0020	
4		4	0.0181	0.1040	29.8568	8.24	0.0043	
5		5	0.0169	0.1209	23.7049	7.81	0.0054	
6		6	0.0122	0.1331	19.8432	5.68	0.0176	
7		7	0.0115	0.1446	16.2819	5.45	0.0201	
8		8	0.0117	0.1563	12.6224	5.61	0.018	

A5727_C

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1		1	0.0247	0.0247	57.1000	11.90	0.0006	
2		2	0.0216	0.0464	47.4722	10.62	0.0012	
3		3	0.0220	0.0684	37.6496	11.03	0.0010	
4		4	0.0152	0.0836	31.4684	7.74	0.0056	
5		5	0.0108	0.0944	27.6705	5.54	0.0190	

ING

Bijlage G: Onderzoek deel IV

Code opstellen betrouwbaarheidsinterval

```
Sub simul()
For i = 1 To Range(100)

        Cells(1 + i, 6) = 0
        Cells(1 + i, 7) = 0
        Cells(1 + i, 8) = 0
       Cells(1 + i, 9) = 0

For j = 1 To Range("total aantal eerste contacten in de eerste n weken voor actie
B+C")
If Rnd < Range(kan eerste  contact op dag1") Then
    Cells(1 + i, 7) = Cells(1 + i, 7) + 1
      End If
      Next

  For j = 1 To Range("N_bc")
If Rnd < Range("kans eerste contact in eerste n weken voor actie B+C) Then
    Cells(1 + i, 6) = Cells(1 + i, 6) + 1
      End If
      Next

 For j = 1 To Range("Grootte_mailing_A")
  If Rnd < Range("kans eerste contact dag 1 actie A") Then
    Cells(1 + i, 8) = Cells(1 + i, 8) + 1

    End If
  Next
  Next
 End Sub
```

ING

Bijlage H: Onderzoek deel IV

Methode

Aanames:

- Dag1=twee dagen na maildatum.
- Eerste n weken van de eerste contacten geven goede benadering totale productie.
- Na n dagen kan al een goede schatting voor het totaal aantal eerste contacten worden gegeven.
- Modelpercentages gemaakt op historische data.

Stap 0. Maak op basis van historische data modelpercentages van wat er gemiddeld per dag binnengekomen is.

Stap 1. Bepaal voor een nieuwe actie het aantal klanten die de eerste n dagen een eerste contacten hebben gemaakt.

Stap2. Bepaal de fractie $p = \frac{aantal\_klanten\_eerste\ ndagen\_nieuwe\_actie}{totale\_mailgroep}$

Stap3. Bouw een BI interval rondom dit aantal door het BI voor fracties toe toepassen.

Stap4. Schat het totale aantal eerste contacten door de volgende formule:

$$Schatting\ totaal\ aantal\ eerste\ contacten = \frac{waargenomen\ aantal\ periode\ 2\ nieuwe\ actie}{model\ percentage\ periode\ 2}$$

Herhaal deze stap ook voor de ondergrens en bovengrens verkregen bij stap3. Op deze manier ontstaat een betrouwbaarheidsinterval voor de verkregen schatting.

Stap 5 .Bepaal door middel van de Postbank conversie percentages de bruto productie die uit deze contacten volgt.

Stap. 6 Bepaald door middel van het bruto/netto percentage de netto productie.

Stap 7. De totale productie van de actie is voor het Postbank kanaal bepaald.