UNIVARIATE STATISTIEK VOOR DE MENSWETENSCHAPPEN

Een Open Leerpakket in R

Univariate statistiek voor de menswetenschappen

Een Open Leerpakket in R

Sven De Maeyer
Jan Ardies
Liesje Coertjens
Dimokritos Kavadias

ACADEMIA
PRESS

P. Van Duyseplein 8
9000 Gent
Tel. 09/233 80 88
info@academiapress.be
www.academiapress.be

Uitgeverij Academia Press maakt deel uit van Lannoo Uitgeverij, de boeken- en multimediadivisie van Uitgeverij Lannoo nv.

Sven De Maeyer, Jan Ardies, Liesje Coertjens en Dimokritos Kavadias
Univariate statistiek voor de menswetenschappen – Een Open Leerpakket in R
Gent, Academia Press, 2013, 260 pp.

ISBN 978 90 382 2133 5
D/2013/4804/107
NUR1 916
U 2005

Inhoudstafel

Ten Geleide 3

Hoofdstuk 1 Wat is statistiek? 9
1.1. Statistiek als 'gevaarlijk' hulpmiddel 11
1.2. Statistiek = ? 13
RESPONSEN 19

Hoofdstuk 2 Het statistisch programma R 23
2.1. R omgeving 25
2.2. R installeren 25
2.3. Installatie van pakketten (*packages*) 29
2.4. Werken met pakketten '*packages*' 30
2.5. Conventies in R 32

Hoofdstuk 3 Data en de datamatrix 35
3.1. Wat is data en wat zijn variabelen? 37
3.2. Het meetniveau van variabelen 38
3.3. De datamatrix 47
RESPONSEN 50

Hoofdstuk 4 Databeheer in R 55
4.1. Soorten data 57
4.2. Het aanmaken van data 59
4.3. Een databestand aanmaken 61
4.4. Werken in en met een dataframe 63
4.5. Bestaande datasets inlezen 68
4.6. Bestaande functies inlezen 70
RESPONSEN 71
Gehanteerde functies 75

Hoofdstuk 5 De frequentieverdeling van een variabele 77
5.1. Absolute en relatieve frequenties 79
5.2. Frequentietabel 81
5.3. Cumulatieve frequenties 82
5.4. Histogram 88
5.5. Grafische voorstellingen van categorische variabelen 92
RESPONSEN 106
Gehanteerde functies 115

Hoofdstuk 6 Parameters van ligging en spreiding 117
6.1. Parameters van ligging 119
6.2. Parameters van spreiding 131
6.3. Grafische weergave van ligging en spreiding: de boxplot 145
RESPONSEN 149
Gehanteerde functies 159

Hoofdstuk 7 Parameters van vorm 161
7.1. Scheefheid 163
7.2. Platheid (Kurtosis) 167
RESPONSEN 170
GEHANTEERDE FUNCTIES 175

Hoofdstuk 8 De (standaard-)normaalverdeling 177
8.1. De normaalverdeling 179
8.2. Z-scores 187
RESPONSEN 196
Gehanteerde functies 200

Hoofdstuk 9 Steekproeftheorie 201
9.1. Wat is een populatie? 203
9.2. Steekproeven 205
9.3. De ene steek is de andere niet 207
9.4. Fouten in steekproeven 213
RESPONSEN 221
Gehanteerde functies 228

Hoofdstuk 10 Inferenties over de verdeling van variabelen in de populatie 229
10.1. Betrouwbaarheidsintervallen rond het gemiddelde 231
10.2. Betrouwbaarheidsintervallen rond de variantie 240
10.3. Betrouwbaarheidsintervallen voor de kengetallen van vorm 246
10.4. Betrouwbaarheidsintervallen voor relatieve frequenties 248
RESPONSEN 250
Gehanteerde functies 260

Ten Geleide

Alvorens dit boek te gebruiken is het belangrijk om een aantal afspraken te maken en toe te lichten hoe dit boek is opgevat en opgebouwd. We doen dit onder de vorm van enkele topics over dit boek.

1. Waarover hebben we het in dit boek?

In wetenschappelijk onderzoek heeft de statistiek een belangrijke plaats ingenomen. Dit geldt ook voor de menswetenschappen. Statistiek is een krachtig instrument dat vaak te pas en te onpas wordt ingezet om wetenschappelijke kennis te onderbouwen uit zeer diverse menswetenschappelijke disciplines: bv. psychologie, sociologie, politieke wetenschappen, pedagogie, taal- en letterkunde,... Dit boek geeft een inleiding in de meest frequent gebruikte concepten en technieken uit de univariate statistiek voor menswetenschappen. Het heeft daarbij tot belangrijkste doel het inzicht van de lezer te verhogen. We starten daarbij van nul en bouwen het statistisch inzicht uit tot het infereren naar een populatie toe van univariate statistieken.

Statistiek maakt vaak gebruik van wiskunde en is eigenlijk een wiskundige toepassing. Dit heeft voor veel mensen tot gevolg dat ze het gebruik van statistiek ontwijken aangezien ze niet goed zijn “met cijfers”. In dit boek proberen we de verschillende concepten en technieken op een niet-wiskundige wijze toe te lichten en te drenken in realistische voorbeelden. In die zin heeft dit boek voornamelijk de toepassing van statistiek voor ogen eerder dan de droge theorie achter de statistiek. Dit neemt niet weg dat we die theorie daar waar nodig niet links laten liggen.

De toepassing van de statistiek staat dus centraal in dit boek. Om statistiek toe te passen bestaan verschillende al dan niet commerciële softwarepaketten (vb. R, SPSS (PASW), SAS, STATA,…). In dit boek maken we gebruik van R. Enkele redenen liggen aan de grondslag van deze keuze. Een eerste reden is het vrij beschikbaar zijn van dit softwarepakket. Het is open-source software die bovendien op verschillende besturingssystemen werkt (Windows, MacOs, Linux). Dit maakt dat dit in letterlijke zin het meest toegankelijke softwarepakket is. Bovendien is het in R mogelijk om zeer diverse analyse technieken toe te passen. De gemeenschap van statistici werkt bijna elke nieuwe techniek uit naar een toepassing in R. Daardoor kunnen de meest recente en gevorderde analysetechnieken vaak

ook toegepast worden in R, wat in andere softwarepakketten niet altijd het geval is.

2. Voor wie is het boek bedoeld?

Dit boek is bedoeld voor zowel studenten als onderzoekers al dan niet uit de academische wereld. Om het boek te kunnen hanteren is geen statistische voorkennis nodig. Het enige wat je onder de knie moet hebben zijn simpele rekenkundige bewerkingen: optellen, aftrekken, vermenigvuldigen, delen, machten en wortels. Voor onderzoekers kan het boek een goede manier zijn om hun kennis op te frissen of bij te benen en om te leren werken met het veerzijdige pakket R.

3. Hoe is het boek ingedeeld?

Het boek kan je grofweg in vijf delen opsplitsen. Het eerste deel maakt je wegwijs in wat we verstaan onder statistiek en laat je ook kennismaken met het softwarepakket dat we zullen hanteren om de statistische analyses uit te voeren (Hoofdstukken 1 en 2). In het tweede deel lichten we toe wat variabelen zijn, de verschillende soorten variabelen en hoe je deze variabelen binnen het pakket R kan beheren (Hoofdstukken 3 en 4). Deel drie reikt je de technieken aan om zicht te krijgen op hoe één bepaald kenmerk verdeeld is binnen de groep van eenheden die je onderzoekt (Hoofdstukken 5, 6, en 7). In een vierde deel van het boek staan we stil bij hoe we vanuit de beschrijvende analyses op steekproefgegevens meer te weten kunnen komen van een de verdeling van een variabele in de hele populatie (Hoofdstukken 8, 9 en 10).

Doorheen het OLP wordt gewerkt met datasets en een bestand dat specifiek geschreven functies voor dit OLP bevat. Deze bestanden zijn te downloaden op de open leeromgeving (Moodle) van Academia Press: http://moodle.academiapress.be.

Op diezelfde plaats kan je tevens aanvullend oefenmateriaal terugvinden.

4. Wat verstaan we onder een “Open Leerpakket”?

De filosofie achter dit boek is dat het je in staat zou moeten stellen om zelfstandig te leren. Tijdens het lezen zullen we vaak beroep doen op jou als lezer om eerst na te denken vooraleer verder te gaan. In die zin heeft

het boek als doel je actief aan het werk te zetten als leerder. Zo wijkt het af van een klassiek boek doordat we met jou als lezer in dialoog gaan.

Al naargelang je voorkennis kunnen delen overbodig of onnodig expliciet overkomen. Door dit boek op een zelfstandige wijze door te nemen kan je zelf bepalen welke delen voor jou belangrijk, interessant, uitdagend,… zijn. Je kiest zelf wat je grondig doorneemt en waar je doorheen wandelt.

5. Wat is de betekenis van de verschillende gebruikte symbolen en lettertypes?

In dit boek maken we gebruik van verschillende symbolen die we telkens bij alinea's zetten. Deze symbolen geven aan wat de lezer mag verwachten in de bijhorende alinea. Hieronder de gebruikte symbolen en hun betekenis.

Een stukje **informatie**, dat je best zeer grondig doorneemt vooraleer verder te gaan.

Een **opdracht om zelf uit te voeren**; achteraan elk hoofdstuk kan je de responslaag voor de opdracht terugvinden.

Voor de analyses maken we gebruik van het softwarepakket R. Alinea's met dit symbool geven aan hoe de behandelde **inhoud toegepast kan worden in het softwarepakket R**.

Naast deze verschillende symbolen maken we ook gebruik van typografie. R wordt aangestuurd door commando's die we kunnen intypen. Deze commando's zullen we altijd weergeven in het 'monospaced' lettertype `courier new`.

R gebruikt ">" om een nieuwe lijn aan te geven waarin je het commando kan ingeven. Wanneer we dus een commando aangeven in dit boek zullen we dit als volgt doen:

```
> mean(Var1)
```

De pijl naar rechts ">" hoeft u echter niet meer als commando te typen in R zelf. In het boek geven we dus weer hoe het eruit zou moeten zien in R, inclusief de ">".

Wanneer R een nieuwe lijn aangeeft, gebeurt dit door het plus-teken "+". Wanneer de commando's te lang worden voor één regel zullen we ook

hiervan gebruik maken analoog met R. Om de leesbaarheid te verhogen zullen er soms extra spaties worden toegevoegd in commando's. Deze hoeft u niet mee over te nemen, ze hebben geen invloed op de werking. Hoofdletters en kleine letters hebben echter wel een invloed.

Andere namen, documenten en menu's staan in *cursief.*

Menu's en keuzes worden als volgt beschreven: *File > Save as*, waarmee we aangeven "kies *'Save as'* uit het keuzemenu *'File'*."

6. Is het belangrijk dat je weet wat er gebeurt bij een statistische analyse?

De volgende "parabel" (met dank aan prof. dr. H. van den Bergh) heeft als belangrijkste boodschap: gebruik enkel statistische technieken indien je weet wat je aan het doen bent. Deze boodschap lijkt simpel. Desalniettemin wagen onderzoekers zich vaak aan het uitvoeren van statistische technieken zonder ze echt te doorgronden. De filosofie van dit boek is hoofdzakelijk je dat inzicht mee te geven. We hopen daarin te slagen….

Een groepje van drie statistici trekt samen naar een statistisch congres in Munchen. Ze verzamelen op het perron van Antwerpen Centraal. Een van hen herkent een collega-onderzoeker en stapt op hem af.
"Wat doe jij hier?" vraagt de statisticus.
"We verzamelen hier met drie onderzoekers die samen naar een congres gaan over kwalitatieve onderzoekstechnieken in Munchen", antwoordt de collega kwalitatief onderzoeker.
"Ah, en hoeveel treinkaartjes hebben jullie gekocht?", vraagt de statisticus.
"Drie uiteraard!", antwoordt z'n collega.
"Hmm, mooi. Wij hebben er ééntje gekocht."
"Hè?! Hoe doen jullie dat dan straks op de trein?", vraagt de andere.
"Oh, eenvoudig. Wij als statistici hebben zo onze methodes.", zegt de statisticus.
Beide groepjes stappen op en nemen dicht bij elkaar plaats. Ergens halfweg is het moment aangebroken. De statistici zien in de verte de controleur afkomen. Als de bliksem verdwijnen ze met z'n drieën en nemen plaats in het toilet. De controleur komt langs, klopt op de deur en vraagt: *"Uw kaartje graag"*. Waarop één van de statistici hun enige

kaartje onder de deur schuift, de controleur het een knipje geeft en vervolgens doorgaat. De kwalitatieve onderzoeker heeft dit geobserveerd en denkt daaruit inzicht te hebben ontwikkeld in hoe dit werkt. Een week later zien ze elkaar opnieuw op het perron in Munchen.

"Hey, we hebben nu ook maar één kaartje gekocht!", zegt de kwalitatieve onderzoeker niet zonder enige trots.

"Oh, mooi." zegt de statisticus. *"Wij hebben er geen gekocht"*.

Helemaal uit z'n lood geslagen vraagt de kwalitatieve onderzoeker: *"Hoe gaan jullie dat doen?"*

"Tja, wij hebben zo onze methodes.", zegt de statisticus.

Ze stappen samen op en nemen dicht bij elkaar plaats. Ergens rond Luxemburg zien ze de controleur in de verte opdagen. Als de bliksem verdwijnen de kwalitatieve onderzoekers samen in de toilet. Waarop de statistici opstaan, en eentje van hen op de deur van de toilet klopt en zegt: *"Ihre Ticket bitte"*. De kwalitatieve onderzoekers schuiven hun ticketje onder de deur, waarop de statisticus het aanneemt en met z'n collega's in de andere toilet kruipt.

HOOFDSTUK 1

Wat is statistiek?

DOELSTELLINGEN:

Na dit hoofdstuk:

- ben je bewust van de gevaren die schuilen in het gebruik van statistiek;
- ken je de verschillende situaties waarbij statistiek kan helpen.

Het woord statistiek is voor vele mensen een gekend woord. Naast het feit dat het de haren doet oprijzen van de gemiddelde persoon, is het zo dat mensen met dit woord vaak naar verschillende zaken verwijzen. In dit hoofdstuk staan we stil bij wat wij precies verstaan onder statistiek, zetten we kort uiteen waarom een goede kennis van statistiek onontbeerlijk is en geven we aan wat de mogelijkheden zijn van statistiek.

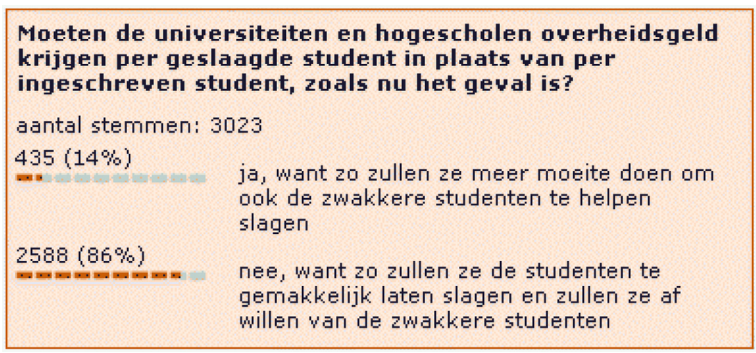
1.1. Statistiek als 'gevaarlijk' hulpmiddel

1.1.1 We starten dit boek met drie voorbeelden van het gebruik van statistiek.

Voorbeeld 1

Op de website van de vrt werd om de andere dag een poll georganiseerd. Hiermee wou de vrt de mening van haar publiek opmeten ten aanzien van diverse topics. Heel vaak zie je dat duizenden mensen deelnemen aan dergelijke polls. De website gaf telkens wel de aantallen weer, maar zegt niets over de waarde van die poll.

Hoe ziet zo'n peiling eruit? De onderstaande illustratie toont een voorbeeld van zo'n poll:



Figuur 1.1: *Voorbeeld van een Poll op de website van VRT*

1.1.2 a) Kan je volgens jou uit de bovenstaande peiling afleiden dat 86% van de Vlamingen vindt dat de universiteiten en hogescholen geen overheidsgeld moeten krijgen per geslaagde student?

b) In juni 2005 heeft de VRT een peiling georganiseerd op haar website over de houding van de Vlamingen ten aanzien van de Europese Grondwet. De poll werd nooit getoond.

Waarom doen ze dat volgens jou? Waarom worden sommige resultaten getoond, terwijl anderen worden geweerd?

1.1.3 Voorbeeld 2

President Bush Jr. van de Verenigde Staten had allerlei wilde plannen met de belastingen voor de Amerikanen vooraleer hij verkozen werd. In het voorjaar van 2003 zei dhr. Bush letterlijk: *"Under this plan, 92 million Americans receive a tax cut of $1083"*. Want volgens zijn berekeningen zou het gemiddelde belastingsvoordeel $1083 bedragen.

1.1.4

Kan je volgens jou uit het feit dat het gemiddelde belastingsvoordeel $1083 bedraagt, afleiden dat de gemiddelde Amerikaan een belastingsvoordeel van $1083 zal hebben onder het nieuwe plan. Wat zou je uit deze woorden kunnen afleiden?

1.1.5 Voorbeeld 3

Een zekere Steven Levitt (econoom) en John Donohue (Jurist) stelden in 2001 dat de legalisering van abortus in de VS vanaf 1973 een invloed had op de mate van criminaliteit, met enige vertraging (nl. 20 jaar)[1].

Op basis van de analyse van misdaadcijfers van 1985 tot 1997, samen met de analyse van abortusgegevens vanaf 1973, bekeken ze patronen van afname in criminaliteit. Deze afname stemt overeen met de periode waarin de kinderen die in de jaren van legalisering zijn geboren, in hun late adolescentie komen. De staten die abortus het eerst legaliseerden, zijn de staten waarin misdaad ook het eerst afnam. De staten met de hoogste mate van abortus zijn ook de staten met de grootste afname in misdaadcijfers.

Het rapport kan je vinden op de CD-Rom in het mapje "Achtergrondliteratuur".

1. Donohue III, J. en Levitt, S. (2001). "The Impact of Legalized Abortion on Crime." *Quarterly Journal of Economics*, 2001, 116(2), pp. 379-420.

1.1.6 Ben je akkoord met de conclusie van de auteurs dat het legaliseren van abortus leidt tot minder misdaad?

1.2. Statistiek = ?

1.2.1 Het woord *statistiek* heeft dezelfde etymologische stam als staat. Als specialiteit werd het ontwikkeld in de periode waarin de moderne natie-staat allerlei instrumenten ontwikkelde om vat te krijgen op de sociale omgeving. In eerste instantie had statistiek te maken met de gegevens die de staat nodig dacht te hebben, gebruikte, om zijn beleid op af te stemmen.

Wat is een staat? Een staat bestaat uit een grondgebied, waarover macht wordt uitgeoefend door allerlei instellingen, en die zijn beslissingen in laatste instantie kan afdwingen door te berusten op geweld. Een staat heeft dus een geweldsmonopolie (of streeft dit alleszins na) en tracht de natuurlijke en sociale omgeving te controleren. Een staat tracht de macht aanvaardbaar te laten zijn door beslissingen te nemen die een zekere waarde hebben voor de onderdanen (of groepen van onderdanen). Eén van de belangrijke instrumenten hierin is het beschikken over informatie en het beheren van sleutelsectoren via informatie.

Moderne staten wilden bijvoorbeeld weten hoeveel mensen er woonden op hun grondgebied. Daarnaast wilden ze weten hoeveel mensen er geboren werden, stierven, verhuisden, waar ze werkten, wat ze deden voor werk, hoeveel geld er in omloop was,… Vanuit de nood naar dit soort informatie is de statistiek ontstaan. Volkstellingen waren de eerste belangrijke statistische instrumenten van een staat.

Daarnaast werden statistieken en statistische technieken ontwikkeld door mensen en organisaties die er winst uit konden slagen. Actuarissen, verzekeringen, banken,... ontwikkelden technieken om geldstromen in kaart te brengen, winstmaximaliserende strategieën te bedenken, levenskansen te berekenen, risico's van verzekerde cargo's in te schatten, ...

1.2.2 Als je zelf het woord statistiek(en) hoort, waaraan denk je dan spontaan?

1.2.3 In welke van de volgende situaties kan statistiek een rol spelen en waarom wel of niet?

a) Om te voorspellen dat een appel naar beneden valt eens hij losraakt van de boom;

b) Om te weten dat kogels door een vitaal orgaan dodelijke verwondingen kunnen veroorzaken;

c) Om te voorspellen dat een kind met een hoog IQ hoge cijfers voor rekenen behaalt op school;

d) Om te weten dat een val van de bovenste verdieping van de Eifeltoren een dodelijke afloop kent.

1.2.4 Statistiek kan drie verschillende functies hebben:

1. Beschrijven
2. Verklaren
3. Voorspellen

1.2.5 ***Beschrijven***

In eerste instantie verzamelden staten eenvoudige gegevens over de bevolking om te weten wie die bevolking is. Statistiek dient dan om een vereenvoudiging te geven van een complexe realiteit, zoals de leeftijdssamenstelling van een populatie.

In 1846 werd, op initiatief van Adolphe Quételet, de eerste volkstelling in België georganiseerd. Hoewel deze tellingen op de eerste plaats een administratieve doelstelling hadden, waren ze van meet af aan ook een belangrijke bron van informatie over de demografische en socio-economische kenmerken van de Belgische bevolking. Anderhalve eeuw later waren tellingen als dusdanig overbodig geworden toen het Rijksregister van de natuurlijke personen de voor de hand liggende bron werd om het bevolkingscijfer te bepalen. Het beschrijft het aantal burgers en dit kan worden gebruikt om bijvoorbeeld de kiesdelers te berekenen. We vergeten deze eerste eenvoudige functie maar al te vaak.

Binnen het domein van de menswetenschappen worden beschrijvingen van een groep mensen gebruikt om meer zicht te krijgen op de eigenschappen van die groep.

1.2.6 

Bedenk zelf een situatie bij een concreet (onderzoeks)probleem in het domein van onderwijs en/of opleidingen waarbij we de beschrijvende functie van statistiek zouden kunnen gebruiken.

1.2.7 ***Verklaren***

Met statistieken kan je een **statistisch model** bouwen. Dit is een grove vereenvoudiging van de realiteit, waarin je beschrijft hoe situaties in gemiddelde termen / in probabilistische termen, werken. Statistiek kan met andere woorden worden ingezet om een bepaald fenomeen dat we vaststellen in de werkelijkheid te verklaren.

Een voorbeeld maakt dit duidelijker.
Er bestaat een verband tussen het roken van tabak en longkanker. Onder de conditie dat alle andere voorwaarden dezelfde zijn (ceteris paribus), heeft iemand die 20 sigaretten per dag rookt 20 keer meer kans om longkanker te krijgen dan een niet-roker. Een meer precies model (meer gedetailleerd model) gaat accurater zijn en bijgevolg ook realiteitsgetrouwer zijn, tegen de prijs van complexiteit. Men kan bijvoorbeeld in het model rekening houden met sekse, leeftijd, passief roken, ander risicogedrag,... Dit model gaat een schatting geven van het aantal rokers dat longkanker zal ontwikkelen in vergelijking tot niet-rokers, en kan gebruikt worden om schattingen te maken over de last voor de sociale zekerheid,...

In dit voorbeeld proberen we het fenomeen “longkanker krijgen” te verklaren. Welke factoren verklaren dat mensen longkankers vormen?

1.2.8 

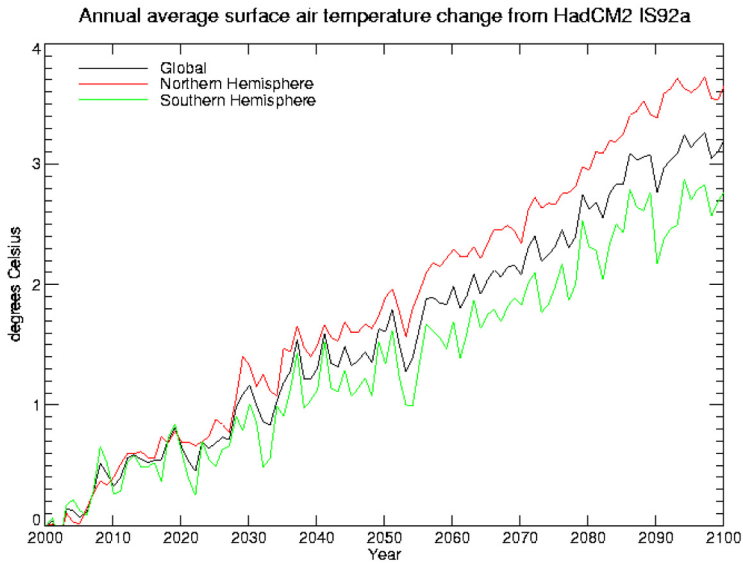
Bedenk zelf een situatie bij een concreet (onderzoeks)probleem in het domein van onderwijs en/of opleidingen waarbij we de verklarende functie van statistiek zouden kunnen gebruiken.

1.2.9 ***Voorspellen***

Een derde mogelijke functie van statistiek is het formuleren van voorspellingen over wat kan gebeuren.

Een voorbeeld van een voorspelling die je kan maken met behulp van statistieken is de gemiddelde temperatuur over een heel jaar in het jaar 2100. De aarde warmt op en de gemiddelde temperaturen zullen tegen 2100

met 3,4° Celsius zijn toegenomen. Statistici hebben deze prognose uitgewerkt en samengevat in de onderstaande grafiek (Fig. 1.2).



Figuur 1.2: *Verwachte opwarming van de aarde voor de periode 2000-2100*

Een ander voorbeeld zijn de levensverwachtingen. De levensverwachting van een man die 36 jaar oud is en die in het Brussels Hoofdstedelijk Gewest woont, bedraagt op dit moment nog een extra 40,85 jaren. Dat kunnen we aflezen uit de onderstaande tabel.

Dit is de statistisch beste gok naar hoeveel jaar een 36-jarige nog voor de boeg heeft. Moest die zelfde persoon in Vlaanderen leven, zou zijn levensverwachting er nog beter uitzien (44,72 jaren).