Statistiek voor Dummies, 2e editie

Spiek Briefje

Enkele belangrijke statistische waarden bepalen

Na het verzamelen van gegevens bestaat de eerste stap uit het berekenen van enkele belangrijke statistische waarden waarmee al direct enkele vragen beantwoord kunnen worden, zoals de volgende:

- Waar bevindt zich het midden van de gegevens?
- Hoe sterk zijn de gegevens gespreid?
- Hoe sterk is het verband tussen de waarden van twee variabelen?

In de volgende tabel vind je de belangrijkste statistische waarden, samen met hun formules en een korte beschrijving.

Statistische waarde	Formule	Toepassing en opmerkingen
Steekproefgemiddelde	$\bar{x} = \frac{\sum x}{n}$	Bepaling van het midden; gevoelig voor uitschieters
Mediaan	Oneven n: middelste waarde van gesorteerde gegevens Even n: gemiddelde van de twee middelste waarden	Bepaling van het midden; niet gevoelig voor uitschieters
Standaarddeviatie van de steekproef	$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$	Maat voor de spreiding; 'gemiddelde' afstand van het gemiddelde
Correlatiecoëfficiënt	$r = \frac{1}{n-1} \sum \frac{(x - \bar{x})(y - \bar{y})}{s_x s_y}$	Sterkte en richting van de correlatie van een lineair verband tussen X en Y

De steekproefomvang statistisch bepalen

De steekproefomvang is het belangrijkste aspect bij het opzetten van een onderzoek, omdat dit de nauwkeurigheid van de gegevens bepaalt. Hoe groter de steekproef, des te nauwkeuriger de gegevens (ervan uitgaand dat je goede gegevens verzamelt) en dus ook het resultaat. Als je weet hoe nauwkeurig jouw resultaten moeten worden (hoe groot de foutmarge mag worden), kun je uitrekenen hoe groot de steekproef moet zijn.

Met de volgende formule bereken je de vereiste steekproefomvang om een bepaald populatiegemiddelde μ te kunnen schatten:

$$n = \left(\frac{z^* \sigma}{FM} \right)^2$$

In deze formule staat FM voor de gewenste foutmarge en is σ de standaarddeviatie van de populatie. Als je deze laatste waarde niet kent, kun je er een schatting van maken door een klein vooronderzoek te houden en de standaarddeviatie van die steekproef (s) te bepalen. Bovendien is z^* de kritieke waarde voor het vereiste betrouwbaarheidsniveau.

Kritieke waarden van betrouwbaarheidsintervallen

Kritieke waarden (z^*-waarden) zijn een belangrijk onderdeel van betrouwbaarheidsintervallen (een statistische methode voor het schatten van populatieparameters). De z^*-waarde die in de formule voor de foutmarge wordt gebruikt, geeft het aantal standaardfouten dat moet worden opgeteld en afgetrokken voor het gewenste betrouwbaarheidsniveau (het percentage zekerheid dat je wilt hebben). Het volgende overzicht geeft enkele gangbare betrouwbaarheidsniveaus met de bijbehorende z^*-waarden.

Betrouwbaarheidsniveau	z^*-waarde
80%	1,28
85%	1,44
90%	1,64
95%	1,96
98%	2,33
99%	2,58

Inhoud in vogelvlucht

Inleiding

Iedere dag krijg je een enorme hoeveelheid statistische informatie voorgeschoteld: diagrammen, grafieken, tabellen en nieuwsberichten over van alles en nog wat, variërend van wetenschappelijk onderzoek tot lezersonderzoeken en sportuitslagen. Dit boek is bedoeld om je op een eenvoudige en begrijpelijke manier wegwijs te maken in deze wirwar van getallen en beweringen. Je leert hoe je statistische informatie over bijvoorbeeld medisch onderzoek moet interpreteren en (heel belangrijk!) hoe misleidend statistieken kunnen zijn. Mocht je zelf op een dag aan de slag willen met statistisch onderzoek, dan laten we zien wat de correcte aanpak is bij het opzetten van het onderzoek, het verzamelen van gegevens, het uitvoeren van berekeningen en het trekken van conclusies.

Dit boek is ook bedoeld als ondersteuning voor degenen die zich voor hun opleiding bezighouden met statistiek. Ik bied niet alleen overzichtelijke informatie en veel duidelijke voorbeelden, maar ook allerlei tips en trucs waarmee het examen een stuk gemakkelijker wordt.

Je vindt hier talloze voorbeelden van statistiek in het dagelijks leven aan de hand van berichtgeving over medische doorbraken, onderzoeken naar criminaliteit en populatietrends, en zelfs een enquête over de slechtste auto's van het millennium! Met dit boek leer je hoe informatie op een correcte en effectieve manier wordt verzameld, weergegeven en geanalyseerd; ook leer je met een kritisch oog te kijken naar de resultaten van onderzoeken, enquêtes, rapporten en experimenten, allemaal zaken waarop vaak belangrijke beslissingen worden gebaseerd. Ik vertel je zelfs hoe je met krekels kunt inschatten hoe warm het is.

Bij dit alles zal ik zo af en toe ook de draak steken met statistici, die dit vak soms veel te serieus nemen. Hoe dan ook, je hoeft helemaal geen hogere opleiding statistiek te volgen om statistische informatie te doorzien.

Over dit boek

Dit boek wijkt op de volgende punten af van traditioneel lesmateriaal over statistiek, zoals de standaardwerken statistiek en studiehandleidingen:

- Het biedt praktische en intuïtieve uitleg van statistische concepten, ideeën, technieken, formules en berekeningen.
- Het laat op een duidelijke, beknopte manier stap voor stap zien hoe je statistische problemen op een intuïtieve manier aanpakt.
- Het biedt interessante praktijkvoorbeelden die voor iedereen in het dagelijks leven herkenbaar zijn.
- Het geeft eerlijk antwoord op vragen zoals 'Wat betekent dit in werkelijkheid?' en 'Wanneer en hoe zal ik dit ooit nodig hebben?'

De conventies in dit boek

Houd de volgende drie conventies goed in gedachten bij het lezen van dit boek:

- **Definitie van de steekproefomvang (*n*).** Wanneer bij een enquête wordt gesproken over de omvang van een steekproef, wordt hiermee het totaal aantal personen bedoeld dat daadwerkelijk heeft deelgenomen aan de enquête en informatie heeft geleverd. Anders gezegd: *n* is de grootte van de uiteindelijke gegevensset.
- **Verwar statistiek niet met statistieken.** Statistieken zijn statistische gegevens, zoals financiële jaaroverzichten, het aantal inwoners per stad enzovoort. Statistiek is de wetenschap die zich bezighoudt met de studie van statistische gegevens, dus ook statistieken, en de verwerking ervan.
- **Gebruik van de term *standaarddeviatie*.** Wanneer de term *standaarddeviatie* (of *standaardafwijking*) wordt gebruikt, gaat het over *s*, de standaarddeviatie in een steekproef. Ik zal het expliciet aangeven als ik de standaarddeviatie van een populatie bedoel.

Dan zijn er nog enkele conventies waarmee ik bepaalde termen en andere tekst aangeef:

- Ik introduceer nieuwe begrippen door ze *cursief* te schrijven.
- Sleutelbegrippen in opsommingen worden aangegeven met **vetgedrukte** tekst.

Wat je kunt overslaan

Ik ben natuurlijk ijdel genoeg om te hopen dat je alles in dit boek de moeite van het lezen waard vindt, maar ik weet ook dat je het al druk genoeg hebt. Je hebt daarom mijn toestemming om desgewenst de uitgebreidere uitleg bij de pictogrammen Technische Info en in de kadertekst (met grijze achtergrond) over te slaan. Deze passages bevatten informatie die interessant kan zijn, maar niet per se nodig is voor jouw begrip van statistiek.

Aannamen

Ik ga ervan uit dat jouw enige ervaring met statistiek tot dusver bestaat uit wat je zoal in de media tegenkomt aan getallen, percentages, tabellen, 'statistisch significante' resultaten, enquêtes, experimenten enzovoort.

Verder neem ik aan dat je enigszins vertrouwd bent met wiskundige basisbewerkingen zoals kwadrateren en worteltrekken, en met wiskundige notatie zoals de letters *x* en *y*, sommatie- en worteltekens enzovoort. Raadpleeg een goed wiskundeboek als dit je helemaal niets zegt.

De opbouw van dit boek

Dit boek bestaat uit vijf delen, elk gewijd aan een belangrijk onderwerp op het gebied van statistiek. Ieder deel bevat meerdere hoofdstukken waarin het hoofdonderwerp op een logische en begrijpelijke manier wordt gepresenteerd. Het zesde deel van dit boek bevat de vertrouwde toptienlijsten, met daarin allerlei interessante, leerzame en soms ook leuke wetenswaardigheden.

Deel I: Onmisbare informatie over statistiek

Dit deel maakt je bewust van de hoeveelheid statistische informatie die je dagelijks over je uitgestort krijgt, en ook welke waarde deze informatie heeft. Je zult zien dat een aanzienlijk deel van de statistische informatie misleidend is, al dan niet met opzet. Je zet je eerste stappen op het glibberige pad van de statistiek door kennis te maken met de achtergronden en methoden die van belang zijn in dit vakgebied. Je leert hoe je statistiek gebruikt als hulpmiddel bij het verkrijgen en interpreteren van informatie; ook maak je kennis met diverse termen uit het statistisch vakjargon.

Deel II: De basis van het rekenwerk

In dit deel raak je vertrouwder met tabellen en grafieken voor het weergeven en interpreteren van verschillende soorten gegevens. Ook lees je hoe je gegevens samenvat en onderzoekt met behulp van bepaalde statistische berekeningen, waarvan je sommige waarschijnlijk al wel kent, maar andere nog niet.

Deel III: Verdelingen en de centrale limietstelling

Dit deel gaat in op de drie meest gangbare statistische verdelingen: de binomiale verdeling, de normale verdeling (inclusief de standaardnormale verdeling oftewel de *Z*-verdeling) en de *t*-verdeling. ik leg uit wat de verschillen zijn tussen deze verdelingen en hoe je iets kunt zeggen over waarschijnlijkheden, percentielen, gemiddelden en de standaarddeviatie. Ook lees je hier het een en ander over de relatieve positie (vergelijkbaar met percentielen).

Als laatste laat ik zien hoe statistici de spreiding van steekproefwaarden bepalen, en wat het belang is van de nauwkeurigheid van deze waarden. Ik eindig met wat wel als een van de belangrijkste basisprincipes van de statistiek wordt beschouwd: de centrale limietstelling.

Deel IV: Betrouwbare schattingen en uitspraken doen

Dit deel behandelt de twee methoden om de resultaten van een steekproef te vertalen in conclusies over de gehele populatie (iets wat ook wel *statistische inferentie* wordt genoemd). Deze twee methoden zijn het betrouwbaarheidsinterval en het toetsen van hypotheses.

In dit deel gebruik je betrouwbaarheidsintervallen om goede schattingen te maken van populatiegemiddelden of -proporties, of van het verschil hiertussen (bijvoorbeeld het gemiddeld aantal uren dat tieners wekelijks voor de televisie hangen, of de verhouding tussen het aantal mannen en vrouwen dat dagelijks medicijnen tegen artritis gebruikt). Ik leg precies uit hoe je een betrouwbaarheidsinterval bepaalt, interpreteert, en controleert op juistheid en geloofwaardigheid. Je leest welke factoren de breedte van een betrouwbaarheidsinterval beïnvloeden (zoals de steekproefomvang), waarbij je leert omgaan met diverse formules, stapvoor-stapberekeningen en voorbeelden van veelgebruikte betrouwbaarheidsintervallen.

Je ziet hoe je een hypothese toetst om aan de hand van jouw gegevens beweringen omtrent gemiddelden of verhoudingen te controleren. Zo beweert een bedrijf dat ze een gemiddelde leveringstermijn van twee dagen hebben, maar klopt dit ook? Je ziet hoe wetenschappers hypotheses (zouden moeten) opstellen en toetsen, en hoe je hun resultaten kunt

controleren op nauwkeurigheid en geloofwaardigheid. Ik leg hierbij stap voor stap aan de hand van voorbeelden uit hoe je de meest gangbare toetsen uitvoert en de resultaten daarvan interpreteert.

Deel V: Statistisch onderzoek en de zoektocht naar relevante verbanden

Dit deel geeft een overzicht van alles op het gebied van enquêtes, experimenten en wetenschappelijke observaties. Je leest wat het doel is van deze onderzoeken, hoe ze worden uitgevoerd, wat de beperkingen zijn en hoe je de geloofwaardigheid van de resultaten inschat.

Verder lees je hoe je paren numerieke en categorische variabelen onderzoekt op verbanden, iets wat bij heel veel onderzoeken gebeurt. Bij paren categorische variabelen maak je kruistabellen en bepaal je de gezamenlijke, voorwaardelijke en marginale kansen en verdelingen. Je controleert de onafhankelijkheid; als je een afhankelijk verband vindt, gebruik je kansberekening om de aard van dit verband te beschrijven. Bij numerieke variabelen gebruik je spreidingsdiagrammen, zoek je naar correlaties, voer je regressieanalyses uit, controleer je de aanpassing van een regressielijn en de invloed van uitschieters, beschrijf je het verband aan de hand van de helling en doe je voorspellingen door middel van de gevonden lijn. En dat allemaal uit de losse pols!

Deel VI: Het deel van de tientallen

Dit deel is klein maar fijn, en doet je onder meer tien manieren aan de hand om snel de betrouwbaarheid van onderzoeken en resultaten te beoordelen. Ook vind je hier tien (legale) manieren om een hoger cijfer te halen bij statistiektoetsen.

Bij sommige statistische berekeningen worden statistische tabellen gebruikt. In de bijlage vind je alle tabellen die je nodig hebt: de *Z*-tabel voor de standaardnormale verdeling (ook wel de *Z*-verdeling genoemd), de *t*-tabel voor de *t*-verdeling, en de binomiale tabel voor (je raadt het al) de binomiale verdeling. Aanwijzingen en voorbeelden voor het gebruik van deze tabellen vind je in de delen van dit boek waarin ze toegepast worden.

De pictogrammen in dit boek

In de marge vind je diverse handige pictogrammen die je op diverse belangrijke dingen wijzen. Dit zijn de pictogrammen en wat ze betekenen:

Dit pictogram kom je tegen bij hints, ideeën en suggesties waarmee je tijd en/of werk kunt besparen. Ook gebruik ik dit pictogram om je een alternatieve kijk op een bepaald onderwerp te bieden.

Dit pictogram geeft belangrijke onderwerpen aan. Onthoud wat je hier leest, want je zult het later in dit boek (en misschien ook daarbuiten) nodig hebben.

Hiermee waarschuw ik voor manieren waarop bijvoorbeeld onderzoekers of de media statistiek kunnen gebruiken om je te misleiden, en wat je eraan kunt doen. Ook geeft dit pictogram belangrijke zaken aan die je op statistiektoetsen en -examens kunt verwachten.

Bij dit pictogram ga ik wat dieper op de materie in dan eigenlijk nodig. Wie het naadje van de kous wil weten, zal deze stukken beslist willen lezen, maar alle anderen kunnen dit met een gerust hart overslaan.

Hoe het verdergaat...

Dit boek is zo geschreven dat je er bijna op iedere willekeurige plaats in kunt duiken en toch kunt volgen wat er gebeurt. Mocht je op zoek zijn naar een bepaald detail over statistiek, zoek het dan op in de inhoudsopgave en begin te lezen op de aangegeven bladzijde. Wil je je verdiepen in een groter deelgebied van de statistiek, begin dan te lezen in het overeenkomstige deel:

- Informatie over grafieken, tabellen, gemiddelden, medianen enzovoort vind je in deel II.
- Mocht je meer willen weten over verdelingen, zoals de *Z*-verdeling, de *t*-verdeling of de binomiale verdeling, ga dan aan de slag met deel III.
- Alles over betrouwbaarheidsintervallen en toetsen van hypotheses lees je in deel IV.
- Voor onderzoeken, enquêtes, experimenten, regressie en kruistabellen moet je in deel V zijn.

Als dit allemaal nog onbekend terrein voor je is, kun je waarschijnlijk het beste beginnen in hoofdstuk 1 en van daaruit de rest van het boek verkennen. Veel leesplezier gewenst!

Hoofdstuk 1

Statistiek in een notendop

In dit hoofdstuk:

- Wat is statistiek nu eigenlijk?
- Het nut van statistiek in jouw dagelijks leven, opleiding en carrière

De wereld waarin we leven wordt niet voor niets de informatiemaatschappij genoemd. Dag in dag uit worden we gebombardeerd met overweldigende hoeveelheden informatie en een groot deel daarvan bestaat uit statistische informatie: 'Onderzoek toont aan dat...' of 'Y procent van de mensen...' of 'Medische doorbraak geeft Z procent hogere overlevingskans...'. Misschien voel je ook regelmatig het nodige wantrouwen wanneer je al dat cijfermateriaal leest. Inderdaad is behoorlijk veel statistische informatie in de media misleidend of ronduit verkeerd, maar het goede nieuws is dat statistiek op nog veel meer manieren positief wordt ingezet en meehelpt onze kwaliteit van leven te verbeteren. Het is slechts een kwestie van de goede van de slechte statistieken te scheiden, zodat je weet wat je kunt vertrouwen en wat niet. De beste manier hiervoor is te leren wat statistiek precies is, en hoe je er op de juiste manier mee omgaat.

In dit hoofdstuk zie je welke rol statistiek speelt in onze van informatie vergeven wereld, en wat je kunt doen om niet alleen in deze wereld te overleven, maar er zelfs je voordeel mee te doen. Ook zul je zien dat statistiek een onmisbaar wetenschappelijk gereedschap is voor het verzamelen van correcte gegevens, het structureren en analyseren van de verzamelde informatie, het interpreteren van de resultaten en het trekken van zinnige conclusies. Statistiek is veel meer dan alleen maar het goochelen met getallen!

Overleven in een wereld van statistiek

Je realiseert je het misschien niet, maar je wordt bijna doorlopend geconfronteerd met statistiek. Wanneer je 's morgens vroeg de krant openslaat voor het weerbericht, krijg je een voorspelling voorgeschoteld die grotendeels is gebaseerd op statistische informatie. De voedingswaarde op de doos ontbijtgranen die je naar binnen knabbelt, bestaat ook uit statistische gegevens. Op je werk maak je misschien offertes die zijn gebaseerd op schattingen van in- en verkoopprijzen, of doe je metingen, of schrijf je rapporten op basis van bepaalde informatie.

Je lunchte voor de verandering in een restaurant dat kortgeleden is uitgeroepen tot de beste gelegenheid in de wijde omtrek. De prijs die je in de supermarkt voor je eten betaalt wordt bepaald door statistische informatie over de productieprijs en de prijs die concurrerende supermarkten willen betalen voor de inkoop dan wel kunnen vragen voor de verkoop.

Als je voor een medische controle naar de dokter gaat, worden dingen zoals de bloeddruk, het gewicht, de temperatuur en allerlei bloedwaarden gemeten en vervolgens vergeleken met statistisch bepaalde waarden.

Je rijdt naar huis in een auto waarvan de boordcomputer gegevens bijhoudt over het brandstofverbruik, de snelheid en de afstand die je rijdt. Wanneer je de televisie aanzet, zie je nieuws over de ontwikkeling van de criminaliteit in jouw woonplaats, maar ook over de aandelenkoersen en (opnieuw) het weer.

Bij het tandenpoetsen gebruik je een tandpasta waarvan statistisch is aangetoond dat deze gaatjes helpt voorkomen, waarna je in bed nog iets leest in een boek uit de (statistisch berekende) lijst met bestsellers. En de volgende dag begint alles opnieuw. Kortom: statistiek speelt een veel grotere rol in je leven dan je zou denken. Hoe weet je nu of je al die statistische informatie kunt vertrouwen? In hoofdstuk 2 lees je meer over de vele manieren waarop statistiek is verweven met ons moderne leven, wat de gevolgen hiervan zijn en hoe je je bewuster wordt van dit alles.

Statistische informatie kan vaag, misplaatst of soms zelf ronduit misleidend zijn. Zorg er allereerst voor dat je statistische informatie in het dagelijks leven als zodanig herkent en dat je niet alles klakkeloos voor waar aanneemt. Wees niet bang om te twijfelen aan wat je tegenkomt; probeer te doorzien wat er bedoeld wordt en ga vragen stellen wanneer dingen verdacht of domweg onbegrijpelijk overkomen. In hoofdstuk 3 leer je hoe statistiek kan worden gebruikt om te liegen, maar ook hoe je deze leugens herkent en doorziet.

Net zoals ieder specialistisch vakgebied heeft ook statistiek zijn eigen jargon. In hoofdstuk 4 maak je kennis met diverse gangbare termen van de statistische vaktaal. Naarmate je het taalgebruik van de statistiek beter leert kennen, wordt het steeds gemakkelijker om de onderliggende principes te vatten. Ook kun je dan veel gemakkelijker begrijpen en uitleggen wat er mis is met bepaalde statistische informatie en waarom. Als je de vaktaal beheerst, wordt het natuurlijk ook een stuk eenvoudiger om zelf statistische informatie te presenteren en toe te lichten. Mocht dit alles je nog steeds niet overtuigen van het nut van statistisch jargon: als je het niet leert, zul je veel meer moeite hebben om dit boek te begrijpen.

In de volgende paragrafen laten we zien hoe statistiek een onmisbaar gereedschap is bij alle stappen van wetenschappelijk verantwoord onderzoek.