Elementaire Statistiek

Elementaire Statistiek

J. van Soest

VSSD

Zevende druk 1992, 1994, 1997
Eerste druk 1972

Uitegegeven door:
VSSD
Leeghwaterstraat 42 2628 CA Delft, The Netherlands
tel. +31 15 27 82124, telefax +31 15 27 87585, e-mail: hlf@vssd.nl
internet: http://www.vssd.nl/hlf
URL met informatie over dit boek: **http://www.vssd.nl/hlf/a013.htm**

Printed in The Netherlands.

ISBN 978-90-407-1270-8
NUR 916
Trefw: statistiek.

Voorwoord

Deze handleiding is geschreven ten behoeve van het college Toegepaste Statistiek, gegeven aan de Technische Universiteit te Delft. Aangezien vele studenten niet aan het vervolgonderwijs toekomen en gelet op de belangstelling die voorgaande uitgaven ondervonden hebben bijvoorbeeld bij het hoger beroepsonderwijs, is gepoogd tot een min of meer afgeronde hoeveelheid basisstof te komen. Het boek bevat dan ook meer onderwerpen dan in het college behandeld worden en een aantal hoofstukken is zeer doelgericht geschreven. Algemene opzet, vele voorbeelden en opgaven zijn afkomstig van de door prof.ir. J.W. Sieben verzorgde colleges en examens aan de TUD.

Ook is er een Aanvulling op Elementaire Statistiek (ISBN 90–6562–006–0) in de handel waarin een belangrijke uitbreiding wordt gegeven van de leerstof in de hoofdstukken 3, 7 en 13.

Bij de auteur is een statistische manipulator verkrijgbaar, te gebruiken op een personal computer als rekenhulp op het gebied van beschrijvende statistiek, keuringen, verdelingsfuncties, toetsen voor aanpassing, betrouwbaarheids-intervallen en eenvoudige regressie- en variantieanalyse.

januari 1992

J. van Soest
Faculteit Technische Wiskunde en Informatica
TU Delft

Inhoud

Inleiding

De naam *statistiek* is ontstaan uit het verzamelen, weergeven en samenvatten van gegevens die nodig waren om de *staat* in stand te houden. Tegenwoordig wordt de statistiek beschouwd als een wetenschap die zich bezig houdt met resultaten verkregen door middel van metingen, enquêtes, enzovoorts. Deze resultaten worden meestal in numerieke vorm gegeven en worden *waarnemingsuitkomsten* genoemd.

Men kan in een statistisch onderzoek drie stadia onderscheiden:

a. het *waarnemen* in de vorm van het *verzamelen* van de gegevens aan de hand van een vraagstelling.
b. het *verwerken* en *presenteren* van de gegevens op beknopte en overzichtelijke wijze. Deze fase staat bekend als de *beschrijvende statistiek*.
c. het *analyseren* en *interpreteren* van de gegevens, hetgeen behoort tot het terrein van de *verklarende statistiek*.

In het laatste stadium trekt men conclusies en neemt men beslissingen op grond van de beschikbare gegevens (die we dan steekproef noemen) omtrent een veel grotere hoeveelheid gelijksoortige gegevens (die met populatie wordt aangeduid). Dit zal door de volgende voorbeelden toegelicht worden.

Voorbeeld 1

Als uit een partij producten een steekproef genomen wordt en deze producten worden op basis van een bepaalde eigenschap als ‘goed’ of ‘slecht’ gekwalificeerd, dan doen we dat om conclusies te trekken over het aantal goede respectievelijk slechte exemplaren in de partij en om te beslissen of deze partij al dan niet voor aflevering geschikt is. □

Voorbeeld 2

Als men gedurende een uur het aantal auto’s telt dat een bepaald punt van een weg passeert, zal men geïnteresseerd zijn in het totale aantal auto’s dat bijvoorbeeld in een jaar langs dat punt komt. De verzameling van alle uurtellingen op dat punt van de weg en in het jaar waarvoor men dat totaal wil weten, vormt de populatie. Deze populatie bestaat dus uit $24 \times 365 = 8760$ uurtellingen, en hieruit is een steekproef van één uurtelling getrokken. □

Voorbeeld 3

Wanneer op een waarnemingspunt van een weg de snelheid van 300 auto's gemeten wordt, zullen we deze snelheidsmetingen opvatten als een steekproef van 300 stuks uit een populatie, bestaande uit de verzameling van alle autosnelheden op dat punt gedurende een bepaalde tijdsperiode. □

Voorbeeld 4

Als de weerstand van een stuk koperdraad een aantal keren met een apparaat gemeten wordt, zal niet steeds dezelfde waarde gevonden worden. Van meting tot meting kunnen allerlei factoren het resultaat beïnvloeden: het inklemmen van de draad zal niet iedere keer op precies dezelfde wijze gebeuren, er kunnen kleine temperatuurvariaties optreden, het aflezen van het apparaat zal niet elke keer even nauwkeurig zijn, etcetera. De meetuitkomsten verschillen dus ten gevolge van toevallige effecten. Wij willen nu uit de verkregen resultaten — de steekproef — een conclusie trekken over de werkelijke weerstand van de draad. Gesteld dat deze resultaten geen systematische afwijking vertonen in die zin dat het verschil tussen het gemiddelde van een zeer groot aantal meetuitkomsten en de werkelijke weerstandswaarde te verwaarlozen is, is de werkelijke weerstandswaarde te vinden als het gemiddelde van een zeer groot aantal meetuitkomsten. De populatie bestaat hier dus uit de resultaten van alle metingen die men onder gelijkblijvende omstandigheden aan de draad zou kunnen verrichten. □

Ook kan een populatie gedefinieerd worden als een verzameling van elementen die voldoen aan een bepaalde omschrijving. Op grond van die omschrijving moet van ieder object kunnen worden vastgesteld of het al dan niet tot de populatie behoort: zo kan men de populatie van bomen in een bos beschouwen als duidelijk aangegeven wordt wat een boom is. Van ieder element van de populatie kan men een bepaalde *eigenschap* waarnemen: als eigenschap van een boom kan de hoogte in aanmerking komen, maar evengoed het aantal bladeren als men daarin geïnteresseerd zou zijn. Wij zullen ons grotendeels beperken tot *kwantitatieve* eigenschappen, ook *variabelen* genoemd. Dat zijn eigenschappen die een numerieke waarde bezitten, en het zal duidelijk zijn dat deze waarde over de elementen van de populatie gezien varieert. Men kan ze onderscheiden in:

a. *continue* variabelen; grootheden die in principe elke waarde in een bepaald interval kunnen aannemen. Bijvoorbeeld de hoogte van bomen.

b. *discrete* variabelen; grootheden waarvoor alleen geïsoleerde waarden in aanmerking komen. Bijvoorbeeld het aantal bladeren van bomen.

Naast kwantitatieve eigenschappen kent men *kwalitatieve* eigenschappen, ook wel *attributen* genoemd. Bijvoorbeeld godsdienst als eigenschap van personen: deze eigenschap kent geen natuurlijke ordening en wordt daarom *nominaal* genoemd. Is er wel een zekere ordening aanwezig, dan spreekt men van een *ordinale* eigenschap. Bijvoorbeeld een indeling van de smaak van biersoorten in goed, matig of slecht. Vaak kan de variabiliteit van zo'n eigenschap met behulp van getallen vastgelegd worden (in voorbeeld 1 via goed = 0 en slecht = 1). Aangezien uitsluitend de eigenschappen van de elementen in de populatie van belang zijn, kan men zonder bezwaar deze eigenschappen zelf als de populatie opvatten; een populatie kan dus beschouwd worden als een verzameling van getallen betreffende een variabele.

Men gaat dus generaliseren van steekproef naar populatie en uiteraard brengt dat het risico met zich mee dat de getrokken conclusie onjuist is. Dit risico zal omschreven worden met behulp van het begrip *kans*. In de statistiek zorgt men er voor dat de kans op het trekken van een onjuiste conclusie (een kans ligt altijd tussen 0 en 1) *klein* is, waardoor men handelt alsof deze fout niet zal voorkomen. Dergelijke risico's zijn in het dagelijks leven heel normaal, zonder die risico's zou men niet kunnen leven. Toch wordt er wel rekening gehouden met gebeurtenissen die een kleine kans bezitten:

a. men verzekert zich tegen dit soort gebeurtenissen indien de gevolgen bij optreden erg ongunstig zijn.
b. men koopt dit soort gebeurtenissen bij een loterij waarbij de gevolgen bij optreden gunstig zijn.

1 Beschrijvende statistiek

1.1. Frequentieverdelingen

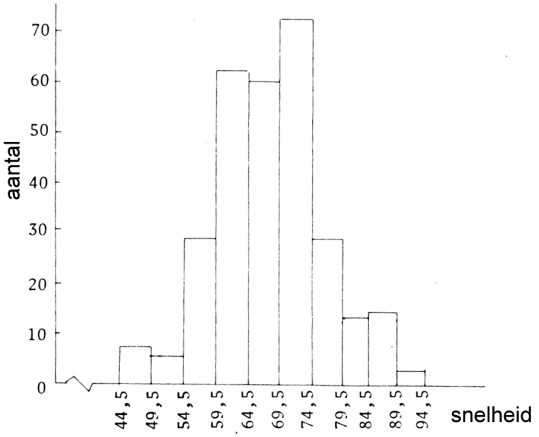
Beschouw een aantal waarnemingsuitkomsten aan een continue variabele. Deze gegevens kunnen overzichtelijk gerangschikt worden door waarnemingen die weinig in grootte van elkaar verschillen, in groepen samen te nemen. Deze groepen worden *klassen* genoemd. Het aantal gegevens in een klasse heet *frequentie*. De som van de frequenties geeft het totale aantal waarnemingen. Men verkrijgt zodoende een *frequentieverdeling*.
In plaats van het aantal waarnemingen in elke klasse kan de fractie of het percentage van het totale aantal waarnemingen aangegeven worden dat in een klasse valt. We krijgen dan een *relatieve frequentieverdeling*.
Ook kan bij elke grens tussen twee klassen aangegeven worden hoe groot het aantal resp. de fractie (percentage) waarnemingen is waarvan de waarde lager is dan die klassegrens. Men spreekt van een *cumulatieve frequentieverdeling* respectievelijk een *relatieve cumulatieve frequentieverdeling*.

Voorbeeld 1.1

Van 300 auto's die op een bepaalde dag een zeker punt van een weg passeerden, is de snelheid bepaald. De resultaten zijn als volgt in een frequentieverdeling en een relatieve frequentieverdeling gegeven:

snelheid in km/uur	aantal auto's	%
klasse	frequentie	relatieve frequentie
45-49	8	2,67
50-54	6	2,00
55-59	29	9,67
60-64	63	21,00
65-69	60	20,00
70-74	74	24,67
75-79	29	9,67
80-84	14	4,67
85-89	15	5,00
90-94	2	0,67
Totaal	300	100,02

Uit de klasse-indeling blijkt dat de autosnelheden bepaald zijn in (dus in feite *afgerond* zijn op) gehele aantallen km/uur. Dit betekent dat bijvoorbeeld de klasse 45-49 alle snelheden van 44,5 tot 49,5 (de *klassegrenzen*) bevat. De *klassebreedte* is dus gelijk aan 5. Onderstaande grafische voorstelling van de frequentieverdeling wordt *histogram* genoemd.



Figuur 1.1.

De cumulatieve frequentieverdeling en de relatieve cumulatieve frequentieverdeling zijn:

Snelheid	aantal	%
< 44,5	0	0,00
< 49,5	8	2,67
< 54,5	14	4,67
< 59,5	43	14,33
< 64,5	106	35,33
< 69,5	166	55,33
< 74,5	240	80,00
< 79,5	269	89,67
< 84,5	283	94,33
< 89,5	298	99,33
< 94,5	300	100,00

□

Bij het maken van een frequentieverdeling en een histogram moeten de volgende regels in acht genomen worden:

a. de klassen moeten zo gekozen zijn, dat het voor iedere waarneming duidelijk is tot welke klasse hij behoort. Daarbij dient rekening te worden gehouden met de wijze waarop de waarnemingen eventueel zijn afgerond.

b. het aantal klassen moet niet te groot zijn om niet te veel onbelangrijke details naar voren te laten komen en anderzijds niet te klein om niet te veel details verloren te laten gaan. Bovendien is het gebruikelijk dat alleen bij een zeer groot aantal waarnemingen meer dan 20 klassen genomen worden en dat het aantal klassen nooit minder dan 5 bedraagt. Verder kan rekening worden gehouden met de eis dat de optimale klassebreedte omgekeerd evenredig is met de derdemachtswortel uit het aantal waarnemingen[1].

c. bij het tekenen van het histogram moet de oppervlakte van de kolommen evenredig zijn met de aantallen waarnemingen die in de betreffende klasse vallen. Alleen als alle klassen even breed zijn, zal dus ook de hoogte van de kolommen evenredig zijn met die aantallen.

Wanneer de waarnemingsuitkomsten betrekking hebben op een discrete variabele, treden in het bovenstaande zekere wijzigingen op; zie voorbeeld 1.2.

Voorbeeld 1.2

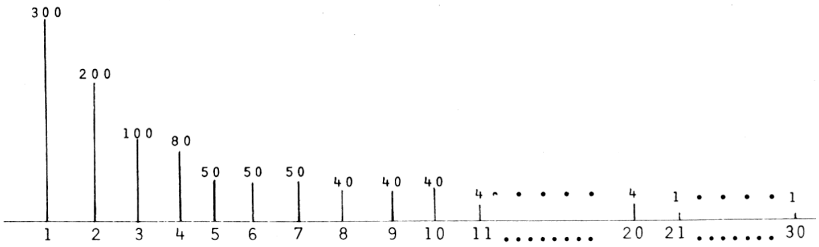
De volgende tabel geeft de frequentieverdeling van 1000 ziektegevallen naar de duur ervan in een groot bedrijf:

duur van de ziekte in dagen	frequentie
1	300
2	200
3	100
4	80
5 t/m 7	150
8 t/m 10	120
11 t/m 20	40
21 t/m 30	10
totaal	1000

De frequentieverdeling van een discrete variabele wordt grafisch voorgesteld als een *staak-* of *staafdiagram*:

---

[1] De evenredigheidsfactor is problematisch: bij een normale verdeling als achterliggend model (zie hoofdstuk 3) is deze $3{,}5\sigma$ waarin σ de standaardafwijking voorstelt. In appendix C geven we enige prenten.



Figuur 1.2. □

1.2. Kentallen voor ligging

1.2.1. Gemiddelden

De algemene ligging van een reeks waarnemingen wordt meestal door een representatief getal (kental) aangegeven in de vorm van een gemiddelde. Als we de waarnemingsuitkomsten aangeven met $x_1, ..., x_n$, dan is een gemiddelde een functie van $x_1, ..., x_n$ die moet voldoen aan de volgende drie eisen:

a. de waarde van de functie mag niet veranderen als de getallen $x_1, ..., x_n$ in een andere volgorde gezet worden. Als men dus drie waarnemingen doet en de getallen 7, 10 en 12 vindt, moet men hetzelfde gemiddelde krijgen als wanneer men deze zelfde getallen bijvoorbeeld in de volgorde 10, 7, 12 zou vinden.

b. als $x_1, ..., x_n$ dezelfde waarde hebben, moet het gemiddelde ook die waarde hebben.

c. als $x_1, ..., x_n$ met eenzelfde bedrag vermenigvuldigd worden en als die nieuwe getallen dan in de functie gesubstitueerd worden, moet de nieuwe waarde van de functie door vermenigvuldiging met datzelfde bedrag uit de oorspronkelijke waarde gevonden kunnen worden.

Er zijn vele functies die aan deze eisen voldoen. De keuze daartussen hangt af van de aard van het cijfermateriaal en van het doel waarvoor men dit materiaal verzameld heeft. Wij noemen:

1) het *rekenkundig gemiddelde*[2]
Deze algemeen bekende grootheid die veel gebruikt zal worden, wordt

---

[2] In plaats van het rekenkundig gemiddelde spreekt men van het 'empirische eerste moment'. Onder het empirische k-de moment wordt verstaan de grootheid $\frac{1}{n}\sum_{i=1}^{n} x_i^k$, terwijl het empirische k-de *centrale* moment gegeven wordt door $\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^k$.

gegeven door

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2) het *meetkundig gemiddelde*

$$x_g = \sqrt[n]{x_1 x_2 \ldots x_n}$$

3) het *harmonisch gemiddelde*

$$x_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

4) het *kwadratisch gemiddelde*

$$x_q = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

Voorbeeld 1.3

Gegeven zijn 5 waarnemingen 7; 9; 12; 13; 14.
Het rekenkundig gemiddelde is

$$\bar{x} = \frac{7 + 9 + 12 + 13 + 14}{5} = 11{,}0$$

Het meetkundig gemiddelde is

$$x_g = \sqrt[5]{7 \cdot 9 \cdot 12 \cdot 13 \cdot 14} = 10{,}7$$

Het harmonisch gemiddelde is

$$x_h = \frac{5}{\frac{1}{7} + \frac{1}{9} + \frac{1}{12} + \frac{1}{13} + \frac{1}{14}} = 10{,}3$$

Het kwadratisch gemiddelde is

$$x_q = \sqrt{\frac{7^2 + 9^2 + 12^2 + 13^2 + 14^2}{5}} = 11{,}3$$

We zien dat $x_h < x_g < \bar{x} < x_q$; deze relatie geldt altijd, tenzij alle (positief geachte) waarnemingen samenvallen. □

Voorbeeld 1.4

Iemand rijdt met een auto een afstand van 2a km. Daarvan wordt a km gereden met een snelheid van s_1 km/u en de andere a km met een snelheid van s_2 km/u. Hij heeft in totaal dus een tijd nodig gehad van

$$\frac{a}{s_1} + \frac{a}{s_2} \text{ uren}$$

en zijn gemiddelde snelheid was

$$\frac{2a}{\frac{a}{s_1} + \frac{a}{s_2}} = \frac{2}{\frac{1}{s_1} + \frac{1}{s_2}} \text{ km/u}$$

Deze gemiddelde snelheid is dus het harmonisch gemiddelde van de afzonderlijke snelheden. Het harmonisch gemiddelde moet voor het gemiddelde van snelheden altijd gebruikt worden als het gaat om het gemiddelde van snelheden over gelijke afstanden. Als het gemiddelde van snelheden gedurende gelijke tijdsperioden gezocht wordt, moet men het rekenkundig gemiddelde gebruiken: als b uren met een snelheid van s_1 km/u en b uren met een snelheid van s_2 km/u wordt gereden, is de afgelegde afstand

$$bs_1 + bs_2 \text{ km}$$

en de gemiddelde snelheid gedurende deze 2b uren dus

$$\frac{s_1 + s_2}{2} \text{ km/u} \qquad \square$$

Opmerking

In dit hoofdstuk wordt met ‘gemiddelde’ steeds bedoeld het rekenkundig gemiddelde.

1.2.2. De mediaan

Het is soms zinvol een andere grootheid, namelijk de *mediaan* x_m, als kental voor ligging van een reeks waarnemingen te gebruiken. De mediaan is de waarneming, die bij rangschikking naar grootte van alle waarnemingen de middelste plaats inneemt[3]. Als het aantal waarnemingen even is, wordt het gemiddelde van de twee middelste waarnemingen genomen. De mediaan maakt wat minder efficiënt gebruik van de aanwezige informatie dan het gemiddelde, maar is makkelijker te bepalen en is *robuust* (dat wil zeggen minder gevoelig)

[3] De mediaan voldoet ook aan de in paragraaf 1.2.1 gestelde eisen.

voor eventuele in het cijfermateriaal voorkomende uitschieters.

Voorbeeld 1.5

Gegeven zijn de waarnemingen: 12; 13; 7; 9; 14.
Naar toenemende grootte gerangschikt wordt de reeks: 7; 9; 12; 13; 14.
De mediaan is dus gelijk aan 12. □

Voorbeeld 1.6

Gegeven zijn de waarnemingen: 12; 12; 13; 9; 7; 14; 5; 4.
Naar toenemende grootte gerangschikt 4; 5; 7; 9; 12; 12; 13; 14.
Dus

$$x_m = \frac{9 + 12}{2} = 10{,}5$$ □

1.3. Kental voor variabiliteit

Onder *variabiliteit*, ook wel *spreiding* genoemd, verstaat men het verschijnsel dat de afzonderlijke waarnemingsuitkomsten onderling verschillen. Hiervoor heeft men als kental de *standaardafwijking* s; deze wordt gegeven door

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}}$$

Het kwadraat van de standaardafwijking heet *variantie*[4].

Voorbeeld 1.7

Van de waarnemingen 12; 13; 9; 7 en 14 is het gemiddelde 11.
De berekening van de standaardafwijking is gegeven in het volgende tabelletje:

---

[4] Een robuust kental wordt gegeven door t = 1,483 MAD, waarin MAD de mediaan van de getallen $|x_i - x_m|$ voor i = 1,2, …, n voorstelt; MAD = Median Absolute Deviation.
De factor n – 1 in s en 1,483 in t maakt dat s^2 en t 'zuiver' zijn indien het waarnemingen aan een normaal verdeelde variabele betreft.