Het berekenen van standaarddeviaties

Eindhoven University of Technology
Computing Centre Note 1980/1

HET BEREKENEN VAN STANDAARDDEVIATIES

J.B. Dijkstra

Het berekenen van standaarddeviaties.

Inleiding.

Beschouw een steekproef van m elementen uit een omvangrijke populatie. De elementen zijn bijvoorbeeld personen en we zouden geïnteresseerd kunnen zijn in hun lengtes. Deze lengtes zijn getallen, aan te duiden als x_i voor i = 1, ..., m. Hierop zijn de volgende drie grootheden gebaseerd:

Gemiddelde : $\bar{x} = \sum_{i=1}^{m} x_i/m$

Variantie : $s^2 = \sum_{i=1}^{m} (x_i - \bar{x})^2/(m - 1)$

Standaarddeviatie: $s = \sqrt{s^2}$

Het nu volgende verhaal gaat over het berekenen van s.

Statistische onzekerheid.

We blijven even bij het voorbeeld van de lengtes van personen. Stel dat we als populatie de bevolking van Nederland beschouwen en de standaarddeviatie binnen deze populatie willen schatten door de standaarddeviatie binnen een steekproef. Er is dan sprake van statistische onzekerheid die zich manifesteert doordat we op grond van de berekende waarde van s^2 een interval kunnen geven voor de populatievariantie σ^2. Dit interval heet een "betrouwbaarheidsinterval" en de grootte ervan hangt nauw samen met de gekozen "onbetrouwbaarheid". Het volgende voorbeeld dient ter toelichting hiervan.

Veronderstel dat de steekproef aselect gekozen is en dat de lengtes van personen normaal verdeeld zijn. Dan geldt het volgende:

$$E\,\underline{s}^2 = \sigma^2$$

$$\frac{\nu\underline{s}^2}{\sigma^2} \approx \underline{\chi}^2_\nu \qquad \text{met } \nu = m - 1$$

De eerste regel zegt dat de verwachting van de steekproefvariantie gelijk is aan de populatievariantie en de tweede regel geeft aan hoe s^2 door toevalligheden in de steekproef kan variëren wanneer σ^2 vaststaat. In woorden staat er: ν maal het quotiënt van de steekproefvariantie en de populatievariantie volgt een χ^2-verdeling met ν vrijheidsgraden. En $\nu+1$ is gelijk aan de steekproefgrootte. Een onderstreping duidt aan dat het om een stochastische variabele gaat.
Stel nu dat s = 20.0 cm en dat dit resultaat gebaseerd is op een steekproef van 30 personen. Er geldt dan $s^2 = 400$ en $\nu = 29$. Een betrouwbaarheidsinterval voor σ^2 met onbetrouwbaarheid $\alpha = 5\%$ kunnen we nu als volgt berekenen. We accepteren een kans van 2½% op waarden kleiner dan de linkergrens en evenzo op waarden groter dan de rechtergrens. In een tabellenboek vinden we:

$$\chi^2_{29}(P = 0.025) = 16.0$$

$$\chi^2_{29}(P = 0.975) = 45.7$$

Substitutie in de zojuist gegeven formule levert de grenswaarden voor σ^2 op. Er geldt:

$$253 < \sigma^2 < 725, \text{ ofwel}$$

$$15.9 < \sigma < 26.9$$

Hierbij is verondersteld dat de x_i's exact bekend waren en dat het rekenproces om tot s te komen zonder fouten is verlopen. De breedte van dit betrouwbaarheidsinterval is uitsluitend bepaald door de statistische onzekerheid. Met behoud van de gekozen onbetrouwbaarheid kan het interval alleen verkleind worden door de steekproefgrootte op te voeren. Het resultaat voor m = 1000 zou bijvoorbeeld zijn:

$$19.2 < \sigma < 20.9$$

In de volgende paragrafen wordt gesproken over het effect op het eindresultaat van meetfouten in de waarnemingen en over numerieke stabiliteit bij het berekenen van s. Het is goed er reeds nu op te wijzen dat kostbare perfectionering op deze twee gebieden niet zinvol is wanneer de statistische onzekerheid relatief groot is.

Onnauwkeurigheden in de waarnemingen.

Laat $X = (x_1, \ldots, x_m)^T$.

Dan geldt voor de Euclidische norm $\| X \|$ van de vector X:

$$\| x \| = (x_1^2 + \ldots + x_m^2)^{\frac{1}{2}}$$

De absolute fout bij het meten van x_i duiden we aan met δ_i. Deze meetfouten vormen samen een vector $\delta = (\delta_1, \ldots, \delta_m)^T$ met als norm:

$$\| \delta \| = (\delta_1^2 + \ldots + \delta_m^2)^{\frac{1}{2}}$$

Terugdenkend aan de lengtes van personen stelt x_i dus de feitelijke lengte voor en $x_i + \delta_i$ het resultaat van de meting. De berekende s, gebaseerd op de waarnemingen, duiden we aan als s_δ. De letter s wordt nu verder gereserveerd voor de standaarddeviatie die zou gelden voor de feitelijke lengtes van de gemeten personen. Rekening houdend met de definitie van s valt eenvoudig in te zien dat

$$s_\delta = s(1 + \Delta) \text{ met}$$

$$|\Delta| \leq \frac{\| \delta \|}{(m-1)^{\frac{1}{2}} s} + o((\frac{\| \delta \|}{s})^2)$$

Vaak komt het voor dat men in staat is een bovengrens voor de relatieve fout te geven. Deze bovengrens γ heeft dan de eigenschap dat:

$$|\delta_i| \leq \gamma |x_i| \text{ voor } i = 1, \ldots, m$$

Substitutie levert in dat geval op:

$$|\Delta| \leq \frac{||X||.\gamma}{(m-1)^{\frac{1}{2}}s}$$

Chan en Lewis (1979) merkten op dat het voor de hand ligt om het conditiegetal K van s te geven als:

$$K = \frac{||X||}{(m-1)^{\frac{1}{2}}s}$$

Er geldt dan dat $K\gamma$ een bovengrens is voor de relatieve fout in s.
Met nadruk zij opgemerkt dat hier geen uitspraak over σ gedaan wordt, en dat de berekening verondersteld wordt foutloos te verlopen.
Eenvoudig valt in te zien dat K alleen groot kan worden als de standaarddeviatie klein is ten opzichte van het gemiddelde. In dat geval is iedere waarneming in benadering gelijk aan dat gemiddelde, zodat geldt:

$$K \approx \frac{\{m.(\bar{x})^2\}^{\frac{1}{2}}}{(m-1)^{\frac{1}{2}}.s}$$

Als nu bovendien de steekproef niet al te klein is, dan kan deze vorm nog vereenvoudigd worden tot $K \approx \bar{x}/s$. Dus geldt voor een redelijke steekproef met x = 1000 en s = 1.00 dat $K \approx 1000$. Als hierbij de metingen met een relatieve precisie van 10^{-4} zijn verricht, dan kunnen we voor s_δ waarden verwachten tussen 0.90 en 1.10. Dit interval kan verkleind worden door de waarnemingen opnieuw te schalen, bijvoorbeeld door er een schatting van het gemiddelde van af te trekken.

Fouten bij de berekening.

De nauwkeurigheid van het eindresultaat hangt natuurlijk ook af van het algoritme en de precisie (woordlengte) waarmee gerekend wordt. Drie verschillende algoritmen zullen hier behandeld worden. Omdat bij alle drie de standaarddeviatie berekend wordt door in de laatste stap de wortel uit de variantie te nemen, zal deze laatste stap in de beschrijving worden weggelaten.

Eerste methode: Definitie.

```
x̄ = 0
for i = 1, 2, ..., m do x̄ = x̄ + x_i od
x̄ = x̄/m
s2 = 0
for i := 1, 2, ..., m do s2 = s2 + (x_i - x̄)^2 od
s2 = s2/(m - 1)
```

Deze methode houdt een directe programmering van de definitie in:

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

Tweede methode: Klassiek.

```
sx = 0
sx2 = 0
for i = 1, 2, ..., m do sx = sx + x_i
                        sx2 = sx2 + x_i^2
                     od
x̄ = sx/m
s2 = (sx2 - sx . x̄)/(m - 1)
```

De werking van deze methode berust op de volgende relatie:

$$\sum_{i=1}^{m} x_i^2 - (\sum_{i=1}^{m} x_i)^2/m = \sum_{i=1}^{m} (x_i - \bar{x})^2$$

Derde methode: West (1979).

```
M = x_1
t = 0
for i = 2, 3, ..., m do q = x_i - M
                        r = q/i
                        M = M + r
                        t = t + (i - 1) * q * r
                     od
s2 = t/(m - 1)
x̄ = M
```

Evenals bij de klassieke methode worden hierbij de waarnemingen slechts één keer beschouwd. De volgende twee formules geven aan hoe het algoritme werkt:

$$m_k = \sum_{i=1}^{k} x_i/k \qquad t_k = \sum_{i=1}^{k} (x_i - m_k)^2$$

Een voordeel van deze methode boven directe programmering van de definitie is dat men op elk moment tussenresultaten kan geven en vervolgens gewoon kan doorrekenen. Ook voor het bijwerken van voorlopige resultaten is deze methode zeer geschikt.

Wilkinson (1963) publiceerde reeds een methode voor analyse van algoritmen bij floating point aritmetiek. Deze analyse is op bovenstaande drie methoden toegepast door Chan en Lewis (1979). Zij kwamen tot het nu volgende resultaat.

Laat η voor een zekere rekenmachine het kleinste getal zijn waarvoor $1 + \eta > 1$. Voor de Burroughs B7700 geldt $\eta \approx 7.28 * 10^{-12}$. $\hat{s}$ stelt nu het getal voor dat de machine aflevert als resultaat bij de berekening van s. Onderstaande formules geven nu een bovengrens voor $|\hat{s} - s|/s$, waarbij als uitgangspunt geldt dat de waarnemingen $x_1, \ldots, x_m$ machinegetallen zijn, en dus exact kunnen worden gerepresenteerd.

Definitie: $2K\eta + (\frac{m}{2} + 1)\eta$

Klassiek : $(\frac{3}{2}m + 1)K^2\eta + \eta$

West : $(\frac{\sqrt{2}}{3}m + 7\sqrt{m} + 1)K\eta + (\frac{m}{2} + 2)\eta$

Het meest opvallende in deze formules is dat de bovengrens voor $|\hat{s} - s|/s$ bij de definitie en het algoritme van West lineair is in het conditiegetal K, terwijl de formule voor de klassieke methode kwadratisch is in K. Bij deze beschouwing over numerieke stabiliteit zijn geen meetfouten verondersteld, en aan het steekproefkarakter van de waarnemingen is geen aandacht besteed. Op het effect van de gesignaleerde verschillen in stabiliteit zal in de volgende paragraaf verder worden ingegaan.

Een simulatie op de B7700.

Op de Burroughs B7700 van het THE-Rekencentrum zijn de drie behandelde methoden toegepast op "steekproeven" van 100 pseudorandom getallen (machinegetallen) uit de normale verdeling met ingesteld gemiddelde $\mu = 1$ en standaardafwijking $\sigma = 1, 10^{-1}, 10^{-2}, \ldots, 10^{-6}$. De getallen zijn gegenereerd volgens de methode van Box en Muller (1958). Zoals reeds eerder is opgemerkt, kan het conditiegetal

$$K = \frac{\| X \|}{\sqrt{m-1} \cdot s}$$

goed benaderd worden door $|\bar{x}|/s$ wanneer $\sigma/|\mu|$ klein is. Bij de aflopende waarden van σ neemt het conditiegetal dus toe.

Onderstaande tabel geeft het aantal correcte cijfers in s^2 bij de drie methoden, toegepast met enkele lengte aritmetiek. Als referentie is een berekening van s^2 volgens de definitie genomen, waarbij dubbele lengte aritmetiek is gebruikt. Op de B7700 geldt hiervoor $\eta \approx 2.65 \times 10^{-23}$, zodat tenminste de eerste 12 cijfers exact zullen zijn (zie de formule voor $|\hat{s} - s|/s$ uit de vorige paragraaf).

σ	definitie	West	klassiek
1	11	10	11
10^{-1}	10	10	7
10^{-2}	11	9	7
10^{-3}	10	9	4
10^{-4}	11	10	2
10^{-5}	11	7	0
10^{-6}	8	6	0

Deze resultaten zijn beter dan men op grond van de vorige paragraaf zou kunnen verwachten. Dat is echter niet verwonderlijk, omdat de formules slechts bovengrenzen voor de relatieve fout aangaven.

Uit de tabel blijkt duidelijk hoe instabiel de klassieke methode is. Het is dan ook betreurenswaardig dat de meeste statistische softwarepakketten deze methode nog hanteren.

Suggesties voor de onderzoeker.

- Als s bedoeld is als schatting voor σ zorg er dan voor dat het effect van meetfouten het steekproefkarakter van de waarnemingen niet overheerst. Anders gezegd: Meet tenminste zo nauwkeurig dat $|s_\delta - s| < |s - \sigma|$.
- Als nauwkeurig meten kostbaar is, heeft het opvoeren van de meetnauwkeurigheid geen zin als $|s_\delta - s| << |s - \sigma|$. Vaak zal het zinvoller zijn om de steekproefgrootte op te voeren.
- Probeer de waarnemingen zodanig te schalen dat σ niet veel kleiner is dan μ.

Suggesties voor de programmeur.

- Als alle waarnemingen in het geheugen kunnen, gebruik dan een directe programmering van de definitie. Dit betreft bijvoorbeeld statistische softwarepakketten die bestemd zijn voor grote machines.
- Als de waarnemingen sequentieel verwerkt moeten worden, gebruik dan de methode van West. Deze is bijzonder geschikt voor "updating", het geval dat de waarnemingen bloksgewijs aangeboden worden en de standaarddeviatie steeds moet worden aangepast. Ook voor zakrekenmachines met weinig geheugen is de methode van West zeer geschikt.
- De klassieke methode is nu opgenomen in veel bestaande statistische programmatuur. Als handhaving hiervan om wat voor reden dan ook noodzakelijk is, laat dan een waarschuwing afdrukken als het conditiegetal K een zekere grens K_0 overschrijdt. Voor de B7700 lijkt $K_0 = 500$ een redelijke keuze.

Literatuur.

1. Tony F. Chan and John Gregg Lewis
Computing Standard Deviations: Accuracy
Communications of the ACM (22), 1979

2. D.H.D. West
Updating Mean and Variance Estimates: An Improved Method
Communications of the ACM (22), 1979

3. J.H. Wilkinson
Rounding Errors in Algebraic Processes
Prentice-Hall, Englewood Cliffs, 1963

4. G.E.P. Box and M.E. Muller
A Note on the Generation of Random Normal Deviates
The Annals of Mathematical Statistics (29), 1958

15 juli 1980.