MATHEMATISCH CENTRUM
AMSTERDAM

Januari 1950

STATISTISCHE AFDELING
Rapport S 33

door

H. Theil en J. Hemelrijk.

Enige opmerkingen omtrent Standaardafwijkingen.

Prof. Borst heeft ons drie vragen gesteld met betrekking tot standaardafwijkingen, die hier achtereenvolgens behandeld zullen worden.

1. Gevraagd de standaardafwijkingen van een enkele bepaling en van het gemiddelde van één duplo-bepaling, wanneer men beschikt over de uitkomsten van 15 tot 50 duplo-bepalingen.

1.1. Stel dat de volgende n duplo-bepalingen zijn verricht:

$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Dan wordt een <u>schatting van de standaardafwijking van één x</u> of van één y gegeven door

$$s_1 = \sqrt{\sum_1^n \frac{(x_i - y_i)^2}{2(n-1)}}$$

Dit geldt voor iedere verdeling van de waarnemingen x en y.

Een <u>restrictie</u> moet echter gemaakt worden: vaak stelt men, dat het interval, begrensd door de gevonden waarde minus twee maal de spreiding enerzijds en door de gevonden waarde plus twee maal de spreiding anderzijds, een massa draagt van 95% 1). Dit nu geldt alleen bij normale verdelingen, zodat deze toepassing van de standaardafwijking bij - vaak voorkomende - niet-normale verdelingen als een veelal vrij grove benadering beschouwd moet worden.

1.2 Thans de standaardafwijking van het gemiddelde van <u>één duplobepaling</u>. Geven we deze weer aan met (x_i, y_i), dan wordt <u>een schatting</u> van de standaardafwijking van één $m_i = \frac{1}{2}(x_i + y_i)$ gegeven door

$$s_2 = \frac{s_1}{\sqrt{2}} = \sqrt{\sum_1^n \frac{\left(\frac{x_i - y_i}{2}\right)^2}{n-1}} .$$

Dit geldt eveneens voor iedere verdeling van x en y, terwijl dezelfde restrictie ten aanzien van het gebruik van s_2 moet worden gemaakt als voor s_1.

---

1) "Een interval draagt een massa van 95%" betekent, dat de waarschijnlijkheid, dat één enkele waarneming in dit interval zal liggen, gelijk is aan 0,95.

. - .

2. Gevraagd de standaardafwijking van het gemiddelde van één duplo-bepaling en van één triplobepaling wanneer men beschikt over de uitkomsten van 20-50 bepalingen uit één monster bloed.

2.1 Stel dat deze uitkomsten zijn

$x_1, x_2, \ldots, x_n$,

en noem het gemiddelde $\bar{x} = \frac{1}{n}(x_1 + \ldots x_n)$. Dan is een schatting van de standaardafwijking van het gemiddelde van één duplobepaling

$$s_3 = \sqrt{\sum_1^n \frac{(x_i - \bar{x})^2}{2(n-1)}}$$

Ook deze schatting geldt voor iedere verdeling ~~bepaling~~ van x, met dezelfde restrictie als die vermeld voor s_1.

2.2 Dit laatste kan bepaald worden voor de formule

$$s_4 = \sqrt{\sum_1^n \frac{(x_i - \bar{x})^2}{3(n-1)}}$$

waarmee de standaardafwijking van het gemiddelde van één triplo- ~~tijds~~bepaling geschat wordt.

3. Gevraagd de standaardafwijking van het product van 2 (of meer) grootheden, waarvan de standaardafwijkingen berekend kunnen worden, en evenzo van het quotient, en tenslotte de standaardafwijking van het gemiddelde van het product (en quotient) als die van de afzonderlijke grootheden bekend zijn.

3.1 In de eerste plaats de standaardafwijking van het product. Stel dat de van de grootheden x en y verrichte waarnemingen zijn:

$x_1, x_2, \ldots, x_n$

$y_1, y_2, \ldots, y_n$,

terwijl de standaardafwijking van $z = xy$ berekend moet worden. Noem het gemiddelde van de x_i resp. van de y_i m_1 resp. m_2 en de standaardafwijkingen resp. s_1' en s_2'. Dus geldt:

$$m_1 = \frac{1}{n} \sum_1^n x_i$$

$$m_2 = \frac{1}{n} \sum_1^n y_i$$

$$s_1' = \sqrt{\sum_1^n \frac{(x_i - m_1)^2}{n-1}}$$

$$s_2' = \sqrt{\sum_1^n \frac{(y_i - m_2)^2}{n-1}}$$

Dan wordt een schatting van de standaarafwijking van $z = xy$ gegeven door

$$s_5 = \sqrt{m_1^2 s_2'^2 + m_2^2 s_1'^2 + s_1'^2 s_2'^2}\,,$$

vooropgesteld, dat de grootheden x en y onafhankelijk van elkaar verdeeld zijn.

Indien de grootheden m_1, m_2, s_1' en s_2' exact gemeten zouden zijn, zou deze formule ook exact gelden, en wel voor iedere verdeling van x en van y. Echter, de 2n grootheden x_i, y_j $(i,j=1,\ldots,n)$ vormen een steekproef, die, indien herhaald, in het algemeen tot andere waarden x_i', y_j' aanleiding zouden geven. Derhalve zullen de uit x_i, y_j berekende grootheden m_1, m_2, s_1', s_2' inhet algemeen afwijken van de "ware" gemiddelden en standaardafwijkingen van de populatie waaruit men de steekproef x_i, y_j getrokken denkt; d.w.z. dat de grootheden m_1, m_2, s_1' en s_2' in het algemeen niet "exact" zullen zijn. De grootheid s_5 moet dan ook als een schatting worden beschouwd.

Hierbij komt nog een tweede moeilijkheid. Ook al zijn de grootheden x en y normaal verdeeld, dan nog is $z = x\,y$ niet normaal verdeeld. Hieruit volgt, dat de stelling, dat het in 1.1 genoemde interval ter breedte van vier maal de standaardafwijking (hier s_5) een massa van 95% draagt, slechts een vrij grove benadering inhoudt.

Ook de voorwaarde van onafhankelijkheid van x en y zal zeker niet altijd vervuld zijn.

3.2 Als men drie grootheden x,y en u met elkaar vermenigvuldigt en van

$$z = xyu$$

de standaardafwijking wil schatten, kan men gebruik maken van de formule

$$s_6 = \sqrt{(m_1^2 + s_1'^2)(m_2^2 + s_2'^2)(m_3^2 + s_3'^2) - m_1^2 m_2^2 m_3^2}\,,$$

waarbij m_1, m_2 en m_3 de gemiddelden van de waargenomen waarden van de variabelen x, y en u voorstellen:

$$m_1 = \frac{1}{n}\sum_1^n x_i\,;\quad m_2 = \frac{1}{n}\sum_1^n y_i\,;\quad m_3 = \frac{1}{n}\sum_1^n u_i$$

en s_1', s_2' en s_3' gegeven worden door

$$s_1' = \sqrt{\frac{\sum_1^n (x_i - m_1)^2}{n-1}}\,;\quad s_2' = \sqrt{\frac{\sum_1^n (y_i - m_2)^2}{n-1}}\,;$$

$$s_3' = \sqrt{\frac{\sum_1^n (u_i - m_3)^2}{n-1}}$$

. - .

MATHEMATISCH CENTRUM
AMSTERDAM

Ook hier is gebruik gemaakt van de hypothese, dat de variabelen x, y en u onafhankelijk van elkaar verdeeld zijn. Voor mmer dan 3 onafhankelijke variabelen kan de formule analoog worden uitgebreid.

3.3 Vervolgens de standaardafwijking van het quotient $q = \frac{x}{y}$ van twee onafhankelijkx verdeelde variabelen x en y.

Met behulp van een reeksontwikkeling 1) kunnen de volgende schattingen worden gegeven:

voor het gemiddelde m van q:

$$m = \frac{m_1}{m_2}\left\{1 + \frac{s_2'^2}{m_2^2}\right\}$$

en voor de standaardafwijking s_7 van q:

$$s_7 = \sqrt{\frac{s_1'^2}{m_2^2} + \frac{3s_1'^2 s_2'^2 + m_1^2 s_2'^2}{m_2^4}}$$

(voor de betekenis van m_1, m_2, s_1' en s_2', zie 3.1)

De hier gegevensschattingen voor gemiddelde en standaardafwijking van 9 kunnen slechts gebruikt worden als de spreiding van y klein is in vergelijking met zijn gemiddelde; als $\frac{s_2'^2}{m_2^2} \leq 0{,}1$ is, is dit wel het geval. Verder geldt ook hier het in 3.1 vermelde bezwaar, dat men hier niet met een normale verdeling te maken heeft, zodat de in 1.1 genoemde stelling slechts met reserve mag worden toegepast . Tevens zal de voorwaarde der onafhankelijkheid van teller en noemer niet steeds vervuld zijn.

3.4 Voor de standaardafwijking van het gemiddelde van een product kan als schatting gebruikt worden

$$s_p = \frac{s_6}{\sqrt{n}}$$

waarbij n het aantal waarnemingen voorstelt. Het spreekt vanzelf, dat voor s_p dezelfde bezwaren gelden als voor s_6 ; is s_6 echter "exact" (zie 3.1), dan is s_p het ook. De reserves ten aanzien van de in 1.1 genoemde stelling verliezen daarmee echter niet hun betekenis.

3.5 Als schatting van de standaardafwijking van het gemiddelde van een quotient kan gebruikt worden

$$s_q = \frac{s_7}{\sqrt{n}} ;$$

verder zij verwezen naar 3.4.

---

1) De volgende reeksontwikkeling is gebruikt:

$$q \approx \frac{x}{m_2} - \frac{x\eta}{m_2^2} + \frac{x\eta}{m_2^3} + \ldots$$

waarbij $\eta = y - m_2$.

4. <u>Product en quotient van niet onafhankelijk verdeelde variabelen.</u>

Indien de variabelen x en y niet onafhankelijk verdeeld zijn, is het niet mogelijk een algemeen verband aan te geven tussen de standaardafwijking van het product (resp.quotient) en de standaardafwijkingen van x en y afzonderlijk. Veelal zullen echter de waarden

$$z_1 = x_1 y_1, \ldots, z_n = x_n y_n \quad \text{resp.} \quad q_1 = \frac{x_1}{y_1}, \ldots, q_n = \frac{x_n}{y_n}$$

wel als onafhankelijke waarnemingen beschouwd kunnen worden. In het bijzonder is dit het geval als de x_i (en eveneens de y_i) onderling onafhankelijk zijn en de afhankelijkheid van x_i en y_i voor iedere i dezelfde is.

In dat geval kan men <u>de standaardafwijkingen van z resp. q</u> schatten door de met de in 3.1 corresponderende formules:

$$s'_z = \sqrt{\sum_1^n \frac{(z_i - \bar{z})^2}{n-1}} \qquad \bar{z} = \frac{1}{n}\sum_1^n z_i$$

resp.

$$s'_q = \sqrt{\sum_1^n \frac{(q_i - \bar{q})^2}{n-1}} \qquad \bar{q} = \frac{1}{n}\sum_1^n q_i$$

De restrictie van punt 1.1 moet ook hier worden gemaakt. De hier beschreven methode is ook toepasbaar, als x en y niet onafhankelijk zijn, maar eist dan, wegens de noodzaak van het berekenen der z_i en q_i, aanzienlijk meer tijd dan de in punt 3.1 en 3.3 beschreven methoden.

5. Uit de voorwaarden en restricties blijkt, dat alle hierboven genoemde formules hun specifieke nadelen hebben. Dit betekent, dat veelal voor bepaalde problemen andere methoden dan die der standaardafwijkingen betere oplossingen geven. Aangezien dit echter afhankelijk is van de bijzonderheden van de betreffende problemen, is het weinig zinvol hierop momenteel nader in tegaan, daar een samenvatting in algemene termen van deze methoden niet kort kan zijn. Het Mathematisch Centrum is echter steeds bereid bij de oplossing van dergelijke problemen behulpzaam te zijn.
