Extra uitleg i.v.m variantie en standaardafwijking van een steekproef

- Gemiddelde:
  Gegeven is een steekproef ttz., een verzameling van n getallen $x_1, x_2, \dots, x_n$.
  Het steekproefgemiddelde is dan het rekenkundig gemiddelde: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  Zijn alle getallen niet verschillend, dan duiden we met $m_1, m_2, \dots, m_k$ de k verschillende waarden aan die optreden met als respectievelijke frequenties (= het aantal keer dat elke waarde voorkomt) $f_1, f_2, \dots, f_k$. Het gemiddelde wordt dan als volgt berekend:
  $\overline{x} = \frac{1}{n}\sum_{j=1}^{k} f_j m_j$, waarbij uiteraard: $n = \sum_{j=1}^{k} f_j$ = totaal aantal getallen
  (Omdat de waarde m_j vermenigvuldigd worden met een getal dat het belang ervan uitdrukt noemt men dit een gewogen som)

- We willen nu een grootheid vinden die uitdrukt hoe de gegeven getallen $x_1, x_2, \dots, x_n$ of $m_1, m_2, \dots, m_k$ gelegen zijn t.o.v. dit gemiddelde. Het gemiddelde ligt ongeveer in het midden maar hoever liggen de andere waarden hiervan af, hoe liggen ze gespreid?

  Hiertoe kunnen we naar de afwijkingen $x_i - \overline{x}$ of $m_j - \overline{x}$ kijken. Maar zoals uit (2.4) blijkt heffen, als je de som maakt, de positieve en de negatieve termen mekaar op. We kunnen dit vermijden door de negatieve afwijkingen positief te maken. Een van de manieren om dit te doen is de afwijkingen kwadrateren.
  We definiëren de steekproef **variantie** s^2 als: $\boxed{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$
  (Waarom delen door $n-1$ en niet n? Zie volgend jaar)

  Als veel waarden ver van het gemiddelde liggen dan zijn voor die waarden de gekwadrateerde afwijkingen groot en dus zal de variantie groot zijn. Liggen alle waarden dichtbij het gemiddelde dan is de gekwadrateerde afwijking klein en dus zal ook de variantie klein zijn. De variantie is dus inderdaad een maat voor de spreiding van de waarden rond het gemiddelde. Door het kwadrateren echter heeft de variantie een andere eenheid dan de grootheden zelf (bvb als de x_i in meter(m), uitgedrukt worden krijgen we m^2 als eenheid voor de variantie), vandaar dat we nog als andere spreidingsmaat de **standaardafwijking** s definiëren: $\boxed{s = \sqrt{s^2}}$. Deze heeft dezelfde eenheid als de waarden zelf en kunnen we dus op dezelfde grafiek voorstellen.

  Gebruiken we de verschillende optredende waarden $m_1, m_2, \dots, m_k$ met hun frequenties $f_1, f_2, \dots, f_k$ dan is de definitie logischerwijze: $\boxed{s^2 = \frac{1}{n-1}\sum_{j=1}^{k} f_j (m_j - \overline{x})^2}$

  Via een korte berekening kunnen we de in de praktijk meer gebruikte uitdrukkingen afleiden: $\boxed{s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)}$ en $\boxed{s^2 = \frac{1}{n-1}\left(\sum_{j=1}^{k} f_j m_j^2 - \frac{1}{n}\left(\sum_{j=1}^{k} f_j m_j\right)^2\right)}$

Of kortweg: $\boxed{s^2 = \frac{1}{n-1}\left(\sum x^2 - \frac{1}{n}\left(\sum x\right)^2\right)}$

Inderdaad:

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

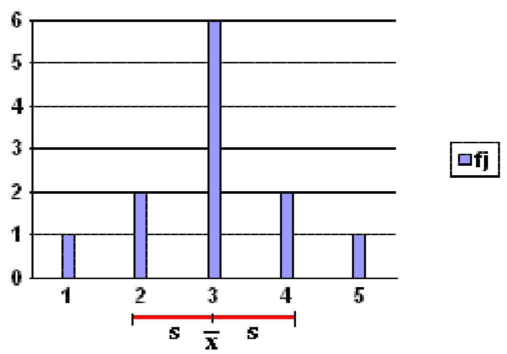
$$= \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \bar{x}^2\sum_{i=1}^{n} 1)$$

$$= \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - 2\frac{1}{n}(\sum_{i=1}^{n} x_i)^2 + n\frac{1}{n^2}(\sum_{i=1}^{n} x_i)^2)$$

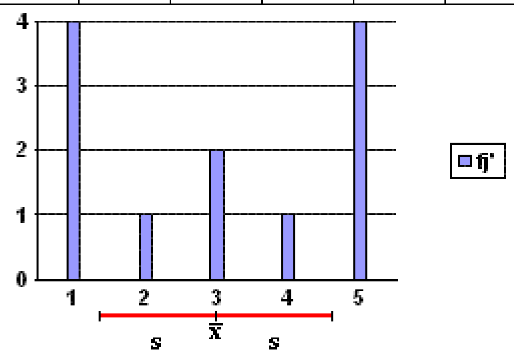
$$= \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2)$$

Een voorbeeld:
Beschouw volgende gegevens: 2x dezelfde waarden maar met verschillende frequenties.
Wat is de invloed van het verschil in frequenties op de variantie en de standaardafwijking?

m_j	1	2	3	4	5
f_j	1	2	6	2	1



m_j	1	2	3	4	5
f_j'	4	1	2	1	4



We zullen de berekeningen maken via de RM.
Kies STAT en dan 1:Edit.
Vul de waarden voor m_j in in kolom L1,

de waarden voor f_j in in kolom L2 en de waarden voor f_j' in kolom L3.

Je vind de waarden voor gemiddelde en standaardafwijking via:
STAT, CALC, 1: 1-Var Stats ENTER kolom waarden, kolom frequenties(indien nodig)

<table>
<tr>
<td>kolommen L1 en L2 nodig:
1-Var Stats L1,L2
Gevolgd door ENTER.
Op het scherm lees je o.a. af:
$\overline{x} = 3$, dit is het gemiddelde
$Sx = 1.044465936$, dit is de standaardafwijking.
De variantie is het kwadraat hiervan:
$\text{var} = 1.0909091 = \frac{12}{11}$</td>
<td>kolommen L1 en L3 nodig:
1-Var Stats L1,L3 ENTER.
Op het scherm lees je o.a. af:
$\overline{x} = 3$, het zelfde van daarnet
$Sx = 1.758098146$, var = 3.0909091
beide groter dan daarnet
Dit is logisch want de waarden die het verst van het gemiddelde liggen hebben hier ook de hoogste frequentie en dragen dus meer bij in de uitdrukking.</td>
</tr>
<tr>
<td colspan="2">Je kan de grootheden die je op je scherm afleest steeds opnieuw opvragen via VARS, 5:Statistics . Willen we bvb de variantie berekenen via de formule:
$s^2 = \frac{1}{n-1}\left(\sum x^2 - \frac{1}{n}\left(\sum x\right)^2\right)$ dan typ je op je RM:
$1 \div 11 \times$(VARS 5: $\sum 2$: $-1 \div 12 \times$VARS 5: $\sum 1$: $\wedge 2$) ENTER
Doe dit voor bovenstaande gevallen….</td>
</tr>
</table>

- Invloed van een verschuiving en een herschaling op gemiddelde, variantie en standaardafwijking.

  In de onderstaande grafiek vind je een illustratie van de transformatie eigenschappen van gemiddelde en variantie/standaardafwijking:
  Als $y = ax + b$ dan is:

$$\overline{y} = a\overline{x} + b$$
$$s_y^2 = a^2 s_x^2$$
$$s_y = |a|\, s_x$$