

Samenvatting Onderzoekstechnieken

TIN 2 - HoGent

Lorenz Verschingel

24 april 2015

1 Het onderzoeksproces

1.1 De wetenschappelijke methode

Aan de hand van **empirisch onderzoek** zijn we geïnteresseerd in volgende zaken:

1. Exploratie
2. Beschrijving
3. Voorspelling
4. Controle

Het onderzoeksproces verloop normaal gezien als volgt:

1. Formuleren
Wat is de onderzoeksvraag
2. Exacte informatie behoefte definiëren
Welke specifieke vragen moeten we stellen
3. Uitvoeren onderzoek
Enquêtes, simulaties. . .
4. Verwerken gegevens
Statistische software
5. Analyseren gegevens
Uitvoeren statistische methodes
6. Conclusie schrijven
Schrijven onderzoeksverslag

1.2 Basisconcepten in onderzoek

1.2.1 Variabelen en waarden

Een variabele is een eigenschap van een object waardoor we objecten van elkaar kunnen onderscheiden.

Een waarde is een specifieke eigenschap, een invullen voor een variabele.

1.2.2 Meetniveaus

De kwalitatieve schalen zijn:

1. Nominaal: Categorieën geslacht, ras, land. . .
2. Ordinaal: Volgorde militaire rang, opleidingsniveau. . .

De kwantitatieve schalen zijn:

1. Interval: Meting: nulpunt is onbelangrijk graden Celsius
2. Ratio: Meting: t.o.v. absoluut nulpunt meter, Joule, kilogram

1.2.3 Verbanden tussen variabelen

Er is een verband tussen variabelen als hun waarde systematisch veranderen.

Men is vooral op zoek naar oorzakelijke verbanden:

- Frustratie leidt tot agressie
- Alcohol leidt tot minder oplettendheid

De oorzaak is hierbij de onafhankelijke variabele.

Het gevolg is de afhankelijke variabele.

Hierbij moet men wel opletten. Een verband tussen variabelen duidt niet noodzakelijk op een oorzakelijk verband.

2 Analyse van 1 variabele

2.1 Beschrijvende statistiek

2.1.1 Centrummaten

Het **gemiddelde** is de som van alle waarden gedeeld door het aantal waarden.

Om de **mediaan** te vinden, sorteert men de waarden en kiest men dan het middelste nummer. Bij een even aantal getallen neemt men het gemiddelde van de twee middelste.

De **modus** is het vaakst voorkomende getal in een reeks getallen. Als men niet onmiddellijk de modus kan aflezen kan men gebruik maken van ranges. Deze ranges zijn dan modale klassen.

2.1.2 Spreidingsmaten

Het **bereik** van een reeks getallen is de absolute waarde van het verschil tussen het grootste en het kleinste getal in de reeks: $|x_{min} - x_{max}|$

De **kwartielen** van een gesorteerde reeks getallen zijn de waarden die de lijst in vier gelijke delen verdeelt. Elk deel vormt dus een kwart van de dataset. Men spreekt van een eerste, tweede en derde kwartiel genoteerd als respectievelijk Q_1 , Q_2 , Q_3 . Hierbij is Q_2 de mediaan.

De **variantie** is het gemiddelde gekwadrateerde verschil tussen de elementen van de dataset en zijn gemiddelde: $\sigma^2 = \frac{1}{n} \sum_i^n (\mu - x_i)^2$

De **standaardafwijking** is hde vierkantswortel van de variantie.

2.2 Eenvoudige grafieken

2.2.1 Cirkeldiagram

Voordelen:

- Met percentages rond 20% kan men makkelijk verduidelijken t.o.v. de volledige dataset.

Nadelen:

- Vergelijking op basis van de hoek.
- De figuur wordt onduidelijk als er veel categorieën zijn.

Men gebruik best zo weinig mogelijk een cirkeldiagram.

2.2.2 Staafdiagram

Voordelen:

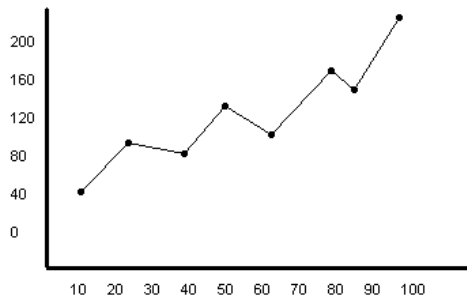
- Categorieën zijn makkelijk te vergelijken.
- Per categorie zijn meerdere staven mogelijk.

2.2.3 Boxplot

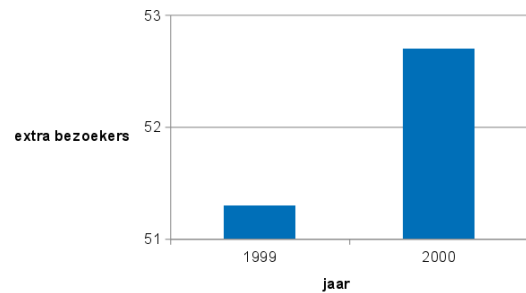
Voordelen:

- Snelle manier om data te inspecteren en verschillende datasets te vergelijken.

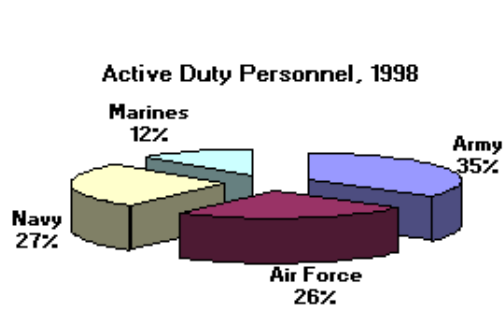
2.3 Interpretatie van grafieken



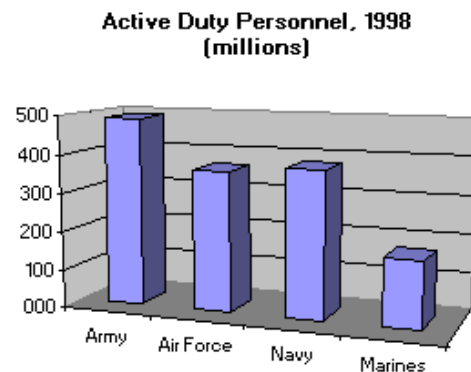
(a) Data-ambigüiteit



(b) Data distortion



(c) Data distraction 1



(d) Data distraction 2

Figuur 1: Valkuilen bij het interpreteren van grafieken

2.3.1 Data-ambigüiteit

Data-ambigüiteit betekent vergeten aan te duiden wat de data betekent. Zie figuur 1a.

Enkele tips om dit te voorkomen:

- Benoem de assen
- Geef een duidelijk titel
- Benoem de meeteenheid (en evt. de grootorde)
- Voeg een bijschrift toe met uitleg over de grafiek

2.3.2 Data distortion

Data distortion betekend dat men verkeerde conclusies kan trekken uit een grafische voorstelling. Zie figuur 1b: merk hierbij op dat de as niet op nul begint en er maar 3 waarden worden weergegeven.

2.3.3 Data distraction

Dit betekent dat de grafiek te veel toeters en bellen bevat. Men moet de *inkt to data ratio* beperken. De figuren 1c en 1d zijn hier voorbeelden van.

3 Analyse van 2 variabelen

3.1 Bivariatie analyse

3.2 Kruistabellen en Cramér's V

Een kruistabel wordt als volgt opgesteld:

1. Percenteer: Deel de cel door het kolomtotaal
2. Bepaal de schatter $e = \frac{\text{kolomtotaal} \times \text{rijtotaal}}{n}$ n: totaal aantal
3. Bepaal het verschil $cel - e$
4. Kwadrateren en normeren $(cel = \frac{\text{verschil}^2}{e})$

$$\chi^2 = \sum \frac{(a-e)^2}{e}$$

$$V = \sqrt{\frac{\chi^2}{n \times (k-1)}} \text{ met: } n = \text{totaal} \text{ en } k = \min(\#rijen, \#kolommen)$$

Cramér's V is een maat die aanduidt hoe sterk de samenhang is tussen twee nominale variabelen. Dit getal ligt altijd tussen 0 en 1.

Waarde	Interpretatie
0	Geen samenhang
0.1	zwakke samenhang
0.25	redelijk sterke samenhang
0.5	sterke samenhang
0.75	zeer sterke samenhang
1	volledige samenhang

Tabel 1: Interpretatie van Cramér's V

3.3 Regressie

Bij regressie gaan we proberen een consistente en systematische koppeling tussen variabelen te vinden.

- **Niet-monotoon:** aanwezigheid (of afwezigheid) van de ene variabele gerelateerd aan de aanwezigheid (of afwezigheid) van een andere variabele.
- **Monotoon:** algemene richting van de samenhang tussen de twee variabelen kan aangeduid worden.

3.3.1 Lineaire regressie

Een lineair verband is een rechte lijnige samenhang tussen een onafhankelijke en afhankelijke variabele, waarbij kennis van de onafhankelijke variabele kennis over de afhankelijke variabele geeft.

KLEINSTE KWADRANTEN METHODE

We proberen een rechte te vinden van de vorm $y = \beta_0 + \beta_1 x$

Hierbij geldt:

$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Kwadraten omdat kleine verschillen minder in rekening moeten gebracht worden dan grote verschillen.

Je kan altijd een kleinste kwadratenmethode uitvoeren. Dit is daarom wel niet altijd goed.

3.4 Correlatiecoëfficiënt en determinatiecoëfficiënt

De Pearson correlatiecoëfficiënt is een maat voor de sterkte van de lineaire samenhang tussen x en y.

De determinatiecoëfficiënt (R^2) verklaart het percentage van de variantie van de waargenomen waarden t.o.v. de regressierechte.

3.4.1 Covariantie

$$\text{cov}(x, y) = \sum_i \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$R = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\text{cov}(x, y)}{\sqrt{\frac{(x - \bar{x})^2}{n} \times \frac{(y - \bar{y})^2}{n}}} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{(\sum (x - \bar{x})^2) \times \sqrt{(\sum (y - \bar{y})^2)}}$$

4 Steekproefonderzoek

4.1 Steekproefonderzoek

De verzameling van alle objecten of personen waar men in geïnteresseert is en onderzoek wil naar doen, heet de *populatie*.

Wanneer met een subgroep uit een populatie gaat onderzoeken, dan noemen we die groep een *steekproef*.

Om tot een steekproef te komen neemt men de volgende stappen:

1. Definitie van de populatie
2. Bepalen van het steekproefkader
3. Budget en tijd

Een *gestratificeerde steekproef* is proportioneel als het aandeel van de subpopulatie in de steekproef gelijk is aan het aandeel van de subpopulatie in de populatie als geheel.

1. **Aselecte steekproef:** elk element uit de onderzoekspopulatie heeft een even grote kans om in de steekproef te komen.
2. **Selecte steekproef:** of een element uit de steekproef terecht komt is afhankelijk van een persoonlijke beoordeling van een onderzoeker.

4.1.1 Fouten bij steekproeven

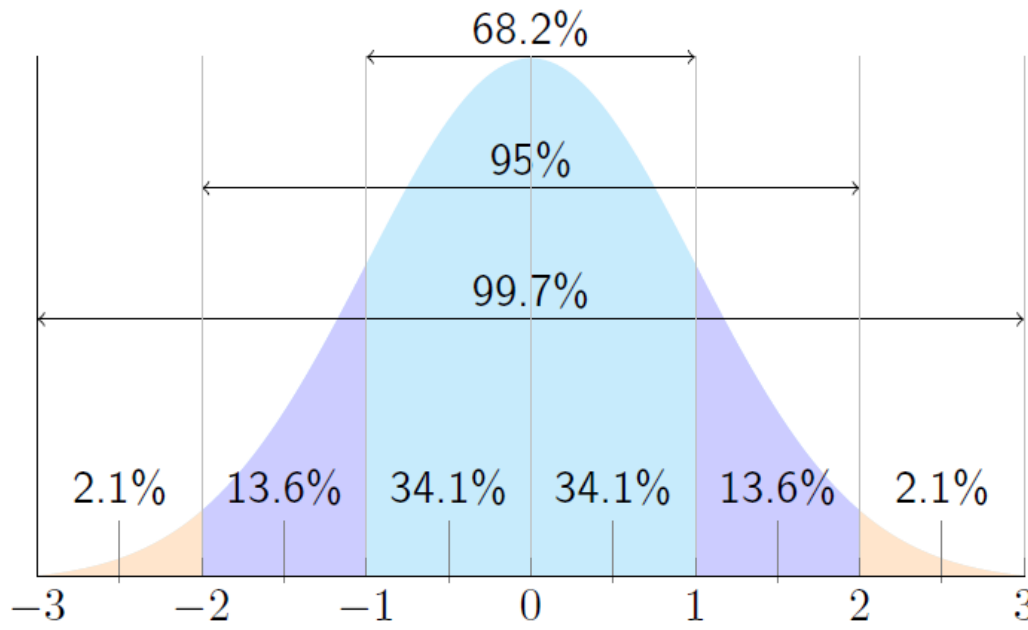
- **Toevallige steekproeffouten:** puur toeval
- **Systematische steekproeffouten:** een fout die een systematische oorzaak heeft.
bv. online enquête sluit een deel van de populatie uit, nl. diegene zonder computer.
- **Toevallige niet-steekproeffouten:** Verkeerd aangekruiste antwoorden.
- **Systematische niet-steekproeffouten:** Respondenten met sterke band met onderwerp van onderzoek reageren positief terwijl anderen niet reageren.

4.1.2 Aanpassing formule standaarddeviatie

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

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

4.2 Kansverdeling van een steekproef



Figuur 2: Standaardnormale verdeling

De formule voor de standaardnormale verdeling, die te zien is op figuur 2, is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \quad (3)$$

Men noteert dat SxS normaal verdeeld is met gemiddelde μ en standaardafwijking σ als: $X \approx Nor(\mu, \sigma)$.

Standaard verdeling: $Z \approx N(\mu = 0, \sigma = 1)$

1. $N(0, 1)$: hiervoor bestaan z-tabellen.
2. Symmetrieregels: $P(Z < -z) = P(Z > z)$
3. 100% kans: $P(Z < z) = 1 - P(Z > z)$

4.3 De centrale limietstelling

Als de steekproefomvang voldoende groot is, dan kan de kansverdeling van het steekproefgemiddelde benaderd worden met een normale verdeling. Dit geldt ongeacht de vorm van de kansverdeling van de individuele waarnemingen.

4.4 Van steekproef naar populatie

4.4.1 Puntchatte

Een puntchatte voor een populatieparameter is een regel of een formule die ons zegt hoe we uit de steekproef een getal moeten berekenen om de populatieparameter te schatten.

4.4.2 Betrouwbaarheidsinterval

Een betrouwbaarheidsinterval is een regel of een formule die ons zegt hoe we uit de steekproef een interval moeten berekenen dat de waarde van de parameter met een bepaalde hoge waarschijnlijkheid bevat.

De betrouwbaarheidscoëfficiënt is de kans dat een willekeurig gekozen betrouwbaarheidsinterval de parameter bevat.

Het symbool voor betrouwbaarheidscoëfficiënt is α .

Voorbeeld: We willen betrouwbaarheidsinterval bepalen waar we 95% dat μ er in ligt.

$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$: Hierbij is μ onbekend.

We zoeken dus z waarvoor geldt dat:

$$P(-z < \mu < +z) = 0.95$$

$$P(-Z < \mu) = 0.025 \text{ en } 0.025 = \frac{\alpha}{2}$$

$$P(Z > \mu) = 0.025$$

$$P(-1.96 > \mu) = 0.025$$

$$P(-1.96 < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1.96)$$

$$P(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}})$$

$$\text{Betrouwbaarheidsinterval: } [\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}]$$

4.4.3 Betrouwbaarheidsinterval voor een kleine steekproef

In plaats van: $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

construeren we: $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

Om een betrouwbaarheidsinterval voor het gemiddelde te bepalen op basis van een kleine steekproef bepalen we:

$$\bar{x} \pm t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right)$$

waarbij $t_{\frac{\alpha}{2}}$ gebaseerd is op $(n - 1)$ vrijheidsgraden. We veronderstellen wel dat we een aselechte steekproef genomen hebben uit een populatie die bij benadering normaal verdeeld is.

4.4.4 Betrouwbaarheidsinterval voor een fractie

$$\bar{p} = \frac{\text{aantalsuccessen}}{n}$$

- Verwachting van kansverdeling van $\bar{p} = p$
- De standaardafwijking van kansverdeling $\bar{p} = \sqrt{\frac{pq}{n}}$
- Voor grote steekproeven is \bar{p} bij benadering normaal verdeeld.

Aangezien \bar{p} een steekproefgemiddelde is van het aantal successen, stelt dit ons in staat een betrouwbaarheidsinterval te berekenen analoog als die voor de intervalschatting van μ voor grote steekproeven.

$$\bar{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{p}\bar{q}}{n}} \quad (4)$$

met $\bar{p} = \frac{x}{n}$ en $\bar{q} = 1 - \bar{p}$

5 Hypothese toetsen

5.1 Toetsen van hypothesen

Een **hypothese** is een idee waarvan nog bewezen moet worden dat het juist is.

Een **hypothesetest** is een controle van een uitspraak over de waarden van één of meerdere populatieparameters.

De **nulhypothese** (H_0) is de hypothese die men probeert te ontkrachten door redenering in het ongerijmde.

De **alternatieve hypothese** (H_1 of H_a) is de hypothese die men wil aantonen.

De **teststatistiek** is de variabele die berekend wordt uit de steekproef.

Het **aanvaardingsgebied** is het gebied van waarden die de nulhypothese ondersteunt.

Het **kritieke of verwerpingsgebied** is het gebied van waarden die de nulhypothese verwerpt.

5.2 Werkwijze

1. Bepalen van de hypothesen (H_0 en H_1)

2. Vastleggen van het significantieniveau (α en n)
3. Toestingsgrootte en waarde ervan bepalen
4. Berekenen en tekenen van het kritieke gebied
5. Conclusies trekken

5.3 Overschrijdingskans

De p-waarde is de kans, indien de nulhypothese waar is, om een waarde te verkrijgen van de toetsingsgrootte die minstens even extreem is als de geobserveerde waarde.

5.4 Kritieke gebied

$$g = \mu - z \times \frac{\sigma}{\sqrt{n}} \quad (5)$$

Bereken $P(M > g) = \alpha \Leftrightarrow P(Z > \frac{g-\mu}{\frac{\sigma}{\sqrt{n}}})$

5.4.1 Tweezijdig toetsen

$$g = \mu \pm z \times \frac{\sigma}{\sqrt{n}} \quad (6)$$

5.5 Samenvatting Z-toets

5.5.1 Overschrijdingskans

$$p = P(Z > \frac{x - \mu_0}{\frac{\sigma}{\sqrt{n}}}) \quad (7)$$

$P < \alpha \Leftrightarrow H_0$ verwerpen.

$P \geq \alpha \Leftrightarrow H_0$ aanvaarden.

5.5.2 Kritieke gebied

$$g = \mu_0 + z \frac{\sigma}{\sqrt{n}} \quad (8)$$

In vergelijking 8 zijn alle waarden buiten z gegeven. Z kan echter makkelijk opgezocht worden in de z -tabel.

Doel	Test op gemiddelde waarde μ van één populatie a.d.h.v één steekproef van n onafhankelijke steekproefwaarden		
Voorwaarde	De populatie is willekeurig verdeeld, n voldoende groot ($n > 30$)		
Type test	Tweezijdig	Eenzijdig links	Eenzijdig rechts
H_0	$\mu = \mu_0$	$\mu = \mu_0$	$\mu = \mu_0$
H_1	$\mu \neq \mu_0$	$\mu < \mu_0$	$\mu > \mu_0$
Verwerpingsgebied	$ z > g$	$z < -g$	$z > g$
Teststatistiek	$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$		

Tabel 2: Samenvatting voor kritieke gebied

5.6 Conclusies en consequentie bij toetsen van hypotheses

	Werkelijke stand van zaken	
Conclusies	H_0 is correct	H_1 is correct
Accepteer H_0	Juist	2 ^e soort fout (β)
Verwerp H_0	Fout van 1 ^e soort	Juist

Tabel 3: Conclusies en consequenties bij toetsen van hypotheses

6 Indexcijfers

6.1 Definities

Een indexcijfer is een getal dat de verandering van een variabele in de loop van de tijd meet in verhouding tot de waarde van de variabele tijdens een bepaalde basisperiode.

$$\text{Indexcijfer I} = \frac{\text{uitkomst}_{\text{datum}}}{\text{uitkomst}_{\text{basis}}} \times 100\%$$

6.2 Enkelvoudige indexcijfers

Enkelvoudige indexcijfers hebben betrekking op slechts één object of artikel.

$$I_p = \frac{P - v}{P_b} \times 100 \quad (9)$$

$$I_q = \frac{Q_v}{Q_b} \times 100 \quad (10)$$

$$I_w = \frac{P_v \times Q_v}{P_b \times Q_b} \times 100 = \frac{I_p \times I_q}{100} \quad (11)$$

6.3 Samengestelde indexcijfers

Een samengesteld indexcijfer is het rekenkundig gemiddelde van een aantal enkelvoudige indexcijfers.

6.3.1 Ongewogen samengestelde indexcijfers

Aan elk element wordt hetzelfde belang gehecht.

$$\bar{I} = \frac{\sum_{i=0}^n I_i}{n} \quad (12)$$

- \bar{I} : ongewogen samengestelde index
- I_i : partiële indexcijfers
- n : aantal partiële indexcijfers

6.3.2 Gewogen samengestelde indexcijfers

Het belang van de onderscheiden partiële indexcijfers is ongelijk: men geeft aan elk partieel indexcijfer een wegingscoëfficiënt.

$$\bar{I} = \frac{\sum_{i=0}^n g_i \times I_i}{\sum_i g_i} \quad (13)$$

- \bar{I} : gewogen samengesteld indexcijfer
- g_i : gewicht of wegingscoëfficiënt van het i^e partiële cijfer
- I_i : partiële indexcijfer

6.3.3 Gewogen gemiddelde van Laspeyres

Het gewogen samengesteld indexcijfer van Laspeyres is het gewogen gemiddelde van partiële indexcijfers waarbij de hoeveelheden in de basisperiode constant gehouden zijn m.a.w. indexcijfer met vaste gewichten of berekening volgens basisjaarmethode.

$$I_W^L = \frac{\sum P_v \times Q_v}{\sum P_b \times Q_b} \quad (14)$$

Het indexcijfer van Laspeyres geeft antwoord op de vraag: “Welke prijs betaalt men vandaag voor een zelfde hoeveelheid van dezelfde goederen?”

6.3.4 Gewogen samengestelde indexcijfer van Paasche

Het gewogen samengesteld indexcijfer van Paasche neemt de hoeveelheden in de verslagperiode als wegingcoëfficiënt m.a.w. indexcijfer met veranderlijke gewichten

$$I_W^P = \frac{\sum P_v \times Q_v}{\sum P_b \times Q_b} \quad (15)$$

6.3.5 Kritiek op Laspeyres en Paasche

Laspeyres' index is een overschatting.

Paasche's index is een onderschatting.

6.3.6 Gewogen indexcijfer van Fisher

Fischer's indexcijfer wordt ook wel eens het ideale indexnummer genoemd voor de volgende redenen:

- Het is een geometrisch gemiddelde wat als een geschikte maat voor gemiddelde van ratio's beschouwd wordt.
- Het neemt zowel de basisjaarhoeveelheden als de huidige hoeveelheden in rekening.
- Er is geen bias.
- Het voldoet aan zowel de time reversal als de factor reversal test.

= het meetkundig gemiddelde van Laspeyres en Paasches indexcijfer.

$$I_W^f = \sqrt{I_W^L \times I_W^P} \quad (16)$$