Statistiek: Spreiding en dispersie

6/12/2013

dr. Brenda Casteleyn

Keu6
Coaching & Onderzoek

1. Theorie

Met spreiding willen we in één getal uitdrukken hoe verspreid de gegevens zijn: in hoeveel verschillende categorieën een reeks gegevens verdeeld zijn of hoe ver het minimum van het maximum verwijderd is. Wanneer we bijvoorbeeld willen weten wat de spreiding aan leeftijd is van een groep mensen kunnen we het verschil tussen de oudste en de jongste nemen, bv. 100 jaar (dit noemen we de variatiebreedte). Dan weten we dat er zowel jonge als heel oude mensen in de groep zijn. Maar als de groep bestaat uit volgende leeftijden (1, 2,5,10,15,30 en 100) is dit getal eigenlijk wat misleidend want de groep bestaat vooral uit jongeren en maar één oudere. Het is daarom beter een cijfer te vinden dat gebaseerd is op alle waarden. Dat is de variantie en standaardafwijking.

1) Spreiding op kwantitatief meetniveau

Variantie: $\frac{1}{N-1}\sum_{i=1}^{k} f_i(x_i-\mu)^2$ (bij ongewogen gegevens is k = N en $f_i = 1$)

Standaardafwijking: $\sqrt{\frac{1}{N-1}\sum_{i=1}^{k} f_i(x_i-\mu)^2}$ (dus de vierkantswortel van de variantie)

Intuïtief kunnen we de standaardafwijking uitleggen als de gemiddelde afwijking ten opzichte van het gemiddelde.

Interpretatie standaardafwijking met behulp van stelling van Chebychev:

Voor om het even welke positieve waarde k geldt : minstens een fractie $1 - \frac{1}{k^2}$ van alle meetwaarden ligt in het interval $]x_{gem}-ks, x_{gem}+ks[$ (zie oefeningen)

Variatiecoëfficiënt (VC) = $\frac{s}{x_{gem}}$ (standaardafwijking delen door gemiddelde). Dit is een relatieve maat (vergelijkbaar met percentages) waardoor we kunnen vergelijken.

2) Spreiding voor ordinale gegevens

Interkwartiele spreidingscoëfficiënt IKS = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$

Ordinale dispersie-index oDi: $1 - \frac{\sum_{i=1}^{k-1}(F_i^* - \frac{1}{2})^2}{\frac{(k-1)}{4}}$ met k = aantal categorieën

(minimumwaarde 0, maximumwaarde 1)

Uitleg: bekijk volgende uiterste mogelijkheden voor spreiding en pas de formule $(F_i^2 - 0{,}5)^2$ toe voor alle relatief gecumuleerde frequenties (dus voor alle categorieën behalve voor de laatste) en maak dan de som.

De helft in de minimumcategorie en de andere helft in de maximumcategorie is het ene uiterste:

	f_i	F_i	F_i^*	$(F_i^*-0,5)^2$
1-5	30	30	0,5	0
6-10	0	30	0,5	0
11-15	0	30	0,5	0
16-20	0	30	0,5	0
21-25	0	30	0,5	0
26-30	30	60	1	
Totaal	60			0

Alles in één enkele categorie is het andere uiterste:

	f_i	F_i	F_i^*	$(F_i^*-0,5)^2$
1-5	0	0	0	0,25
6-10	0	0	0	0,25
11-15	0	0	0	0,25
16-20	60	10	1,0	0,25
21-25	0	10	1,0	0,25
26-30	0	10	1,0	
Totaal	60			1,25

Als alles in één categorie zit, krijgen we voor elke categorie 0,25 voor $(F_i^2 - 0,5)^2$. We kunnen de uitkomst hiervan dan veralgemenen naar (k-1)keer 0,25 of (k-1)* 1/4

We willen nu een genormeerde formule krijgen, dwz een formule waarbij het maximum 1 is en overeenkomt met minimale spreiding (dus alles in 1 categorie) en minimum =0 bij maximale spreiding .

Wanneer we $\sum_{i=1}^{k-1}(F_i^* - 0,5)^2$ delen door (k-1)/4 krijgen we 1 voor een minimale spreiding (omdat de teller dan gelijk is aan de noemer) en 0 voor maximale spreiding (omdat de teller dan = 0) en alle andere waarden zitten tussen 0 en 1. Omdat we de maximale spreiding gelijk willen stellen aan 1 en de minimale aan 0 gebruik we als formule het complement nl:

$$1 - \frac{\sum_{i=1}^{k-1}(F_i^* - 0,5)^2}{(k-1)/4}$$

<u>3) Spreiding voor nominale gegevens</u>

Nominale dispersie-index nDi: $\frac{1-\sum_{i=1}^{k}(F_i^*)^2}{\frac{(k-1)}{k}}$ met k = aantal categorieën

(minimumwaarde 0, maximumwaarde 1)

Entropie: het gewogen gemiddelde van de onzekerheden die bij de verschillende categorieën horen. Wegenscoëfficiënten zijn de relatieve frequenties f_i^*

Entropie H = $\frac{-\sum_{i=1}^{k}(f_i^*.lnf_i^*)}{ln2}$ (= absoluut)

Entropie H = $\frac{-\sum_{i=1}^{k}(f_i^* . ln f_i^*)}{ln\, k}$ (= relatief)

Effect van lineaire transformatie op centrum- en spreidingsmaten:

Bij vermenigvuldigding van elke waarneming met een positief getal b (bv. wanneer we van euro naar Belgische frank gaan) worden ook de centrummaten (gemiddelde en mediaan) en de spreidingsmaten (interkwartielafstand en standaardafwijking) met b vermenigvuldigd.

Bij het optellen van hetzelfde getal a (positief of negatief) bij elke waarneming wordt a opgeteld bij de centrummaten en kwartielen en andere percentielen, maar worden de spreidingsmaten niet veranderd.

Lineaire interpolatie

Met lineaire interpolatie kunnen we de meetwaarde van de r-de waarneming schatten binnen een bepaalde klasse i:

$$\tilde{x}_r = l_i^e + \frac{r - F_{i-1}}{f_i} . v_i$$

met i = klassenummer
$\tilde{x}_r$ =lineair geïnterpoleerde meetwaarde van de r-de score
l_i^e = exacte benedengrens van de i-de klasse
r = het rangnummer of de gecumuleerde frequentie van de score
F_{i-1} = gecumuleerde frequentie van de klasse voorafgaand aan de i-de klasse
f_i = absolute frequentie van de i-de klasse
v_i = klassebreedte van de i-de klasse

Voor de berekening van geïnterpoleerde mediaan gebruik je dezelfde formule maar schrijf ipv r de formule van Mediaan nl (N+1)/2 en vervang i door Me

Analoog voor kwartielen.

2. Oefeningen

Examen Tijssen januari 2007 Vraag 1

Stel de eerste klasse gaat van 1-10, de tweede van 11-20, de derde van 21-30, de vierde en laatste klasse van 31-40. Verder is N = 50, f_2 = 20; F_3^* = 0,8 en de frequentiedichtheid van de eerste klasse is 1.

a) Bepaal het verschil tussen het lineair geïnterpoleerde negentigste en tiende percentiel als je ervan uitgaat dat de betrokken variabele weliswaar discreet gemeten is maar latent toch continu is.

b) Wat leert je de voorgaande uitkomst?

c) Bereken op basis van de stelling van Chebychev de grenzen waartussen zich ten minste 80% van de waarnemingen in de bovenvermelde frequentieverdeling zou moeten bevinden?

d) Hoeveel procent van de waarnemingen bevindt er zich daadwerkelijk tussen de grenswaarden uit onderdeel c? Verklaar het verschil.

Examen Tijssen januari 2007 Vraag 3

Verklaar de noemer in de formule van de oDi

Examen Tijssen januari 2008 Vraag 1

In de onderstaande tabel wordt een geclassificeerd overzicht gegeven van de schaalscores voor de variabele 'utilitair individualisme' op basis van een enkelvoudig aselecte steekproef. Hoewel de ruwe schaalscores gehele waardenzijn, nemen we wel aan dat deze variabele latent toch continu is en kwantitatief gemeten is.

	f_i	F_i	F_i^*	Frequentie-dichtheid
1-5	12			
6-10				
11-15	33			
16-20	45	107		
21-25				
Totaal			1,00	24

a) Vul de frequenties in de bovenstaande tabel verder aan.

b) Bereken de lineair geïnterpoleerde interkwartiele spreidingscoëfficiënt

c) Tussen welke grenzen zal volgens de stelling van Chebychev ten minste 75% van de waarnemingen zich bevinden?

d) Hoeveel procent van de waarnemingen ligt feitelijk tussen deze grenzen?

e) Hoe verklaar je het verschil?

Examen Thijssen januari 2008 Vraag 3

Na een lineaire transformatie van de vorm Y = 2X + 20 is de standaardafwijking Y = 10 en gemiddelde Y = 100. Wat is de variatiecoëfficiënt van de oorspronkelijke variabele X?

A. 0.125
B. 8
C. 0.1
D. 10

Examen Thijssen januari 2012 Vraag 1

Gegeven:

	f_i	F_i	F_i^*	Frequentie-dichtheid	Klassencentra
1-10				1,5	
11-20	50				
21-30			0.80		
31-40					
Totaal	100				

a) Bepaal het verschil tussen het lineair geïnterpoleerde negentigste en tiende percentiel.

b) Wat leer je uit voorgaande uitkomst?

c) Bereken op basis van de stelling van Chebychev de grenzen waartussen zich tenminste 70% van de waarnemingen bevinden.

d) Teken het ogief en bepaal grafisch de mediaan

3. Oplossingen

Examen Tijssen januari 2007 Vraag 1

Gegeven: Stel de eerste klasse gaat van 1-10, de tweede van 11-20, de derde van 21-30, de vierde en laatste klasse van 31-40. Verder is N = 50, $f_2 = 20$; $F_3^* = 0{,}8$ en de frequentiedichtheid van de eerste klasse is 1.

	f_i	F_i	F_i^*	Frequentie-dichtheid
1-10				1
11-20	20			
21-30			0,8	
31-40				
Totaal	50			

Gevraagd:

a) Bepaal het verschil tussen het lineair geïnterpoleerde negentigste en tiende percentiel als je ervan uitgaat dat de betrokken variabele weliswaar discreet gemeten is maar latent toch continu is.

b) Wat leert je de voorgaande uitkomst?

c) Bereken op basis van de stelling van Chebychev de grenzen waartussen zich ten minste 80% van de waarnemingen in de bovenvermelde frequentieverdeling zou moeten bevinden?

d) Hoeveel procent van de waarnemingen bevindt er zich daadwerkelijk tussen de grenswaarden uit onderdeel c? Verklaar het verschil.

Oplossing: vul de ontbrekende waarden in in de tabel. Begin met het percentage te berekenen voor de tweede klasse: 20/50 = 0,4. Uit de frequentiedichtheid van de eerste klasse bereken je f. De frequentiedichtheid is immers het absolute aantal/klassebreedte. De klassebreedte is 10, dus het absolute aantal voor de eerste klasse is dan = 10. Nu kun je ook daarvan het percentage berekenen: 10/50 = 0,2. Nu kun je de gecumuleerde percentages invullen. Vermits de tweede gecumuleerde frequentie gelijk is aan 0,6 en de derde aan 0,8 weet je ook dat het percentage van de derde klasse 0,2 is.

	f_i	F_i	F_i^*	Frequentie-dichtheid
1-10	**10**	**10**	**0,2**	1
11-20	20	**30**	**0,6**	**2**
21-30	**10**	**40**	0,8	**1**
31-40	**10**	**50**	**1**	**1**
Totaal	50			**5**

a) Voor verschil tussen het lineair geïnterpoleerde negentigste en tiende percentiel gebruik je volgende formule:

$$\tilde{x}_r = l_i^e + \frac{r - F_{i-1}}{f_i} . v_i$$

en je vervangt r door de formule voor percentiel en i door P_{90} en P_{10}. De formule voor 90ste percentiel is $P_{90} = 90(N+1)/100$ en voor 10de = $P_{10} = 10(N+1)/100$. Uit deze formule weten we dat het 90ste percentiel zich bevindt op de 45,90 ste plaats en het 10de op de 5,1de plaats. Dat is in de vierde en eerste klasse.

$\tilde{P}_{90} = l_i^e + \frac{90(N+1)/100 - F_{P_{90}-1}}{f_{P_{90}}} . v$ (v is klassebreedte van 90ste percentiel)

en

$\tilde{P}_{10} = l_i^e + \frac{10(N+1)/100 - F_{P_{10}-1}}{f_{P_{10}}} . v$ (v is klassebreedte van 10de percentiel)

Bepaal de exacte grenzen: deze zitten exact tussen het einde van klasse 1 en het begin van klasse 2. Vermits er 1 tussen zit, is de helft ervan 0,5. De exacte bovenklasse van de eerste klasse is dus 10+0,5 = 10,5. Om de benedengrens te vinden trekken we een half af van 1. De exacte ondergrens van de eerste klasse is dus 1-0,5 = 0,5. Bij de vierde klasse is de exacte ondergrens 30,5. Vervolgens bepalen we de gecumuleerde frequentie van de voorgaande klasse: bij de eerste klasse is dat 0 en bij de laatste klasse is dat 40. Verder hebben we de absolute frequenties nodig van klasse 1 en 4: dat is telkens 10. De klassenbreedte is 10. We kunnen nu alle waarden in de formules invullen:

$\tilde{P}_{90} = 30,5 + \frac{45,90 - 40}{10} . 10 = 36,4$

$\tilde{P}_{10} = 0,5 + \frac{5,1 - 0}{10} .10 = 5,6$

Het verschil is dus 36,4-5,6 = 30,8

b) Wat leert je de voorgaande uitkomst?

80% van de waarnemingen bevinden zich tussen 5,6 en 30,8.

c) Voor om het even welke positieve waarde k geldt : minstens een fractie 1 - $\frac{1}{k^2}$ van alle meetwaarden ligt in het interval $]x_{gem}-ks, x_{gem}+ks[$

Hiervoor moeten we de standaardafwijking s berekenen en k

Gevraagd is om als fractie 80% te nemen, dus $1 - \frac{1}{k^2} = 0,80$

Hieruit kunnen we k berekenen. --> $- 1/k^2 = 0,80 - 1$

--> $1/k^2 = 0,20$

$$\rightarrow k^2 = 1/0{,}20$$

$$\rightarrow k^2 = 5 \text{ dus } k = 2{,}24$$

Dus minstens 80% van de waarnemingen ligt tussen x_{gem}-2,24.s en x_{gem} +2,24.s

We berekenen nu het gemiddelde en de standaardafwijking.

Voor het gemiddelde maken we de som van $x_i * f_i$ en voor x_i gebruiken we telkens het klassemidden. Deze som is 975. Deze delen we door N om het gemiddelde te bekomen: 975/50 = 19,5.

Om de standaardafwijking te bekomen maken we eerst de verschillen ten opzichte van het gemiddelde. Deze kwadrateren we en vervolgens vermenigvuldigen we deze kwadraten met de frequenties. Ten slotte maken we de som: deze is dan gelijk aan 5200. Om de standaardafwijking te bekomen delen we deze som door N-1 en vervolgens nemen we de vierkantswortel: $\sqrt{\frac{5200}{49}} = 10.30$

	f_i	Klassemidden (x_i)	$x_i * f_i$	$x_i - x_{gem}$ (verschil tov gemiddelden)	$(x_i - x_{gem})^2$ (Kwadraten van verschil)	$f_i(x_i - x_{gem})^2$
1-10	10	5,5	55	-14	196	1960
11-20	20	15,5	310	-4	16	320
21-30	10	25,5	255	6	36	360
31-40	10	35,5	355	16	256	2560
Totaal	50		975			5200

$=\sum f_i(x_i - x_{gem})^2$

Toepassing van de formule Van Chebychev:

80% ligt tussen x_{gem}-2,24.s en x_{gem} +2,24.s

Dus 80% ligt tussen 19,5 - 2,24*10,30 en 19,5 + 2,24*10,30

80% ligt tussen -3,6 en 42,6

d) We stellen vast dat alle waarden (100%) zich tussen deze grenzen bevinden. De ondergrens ligt immers lager dan de laagste grens (-3,6 < 1) en de bovengrens ligt hoger dan de hoogste grens (42,6 > 40).

<u>Examen Tijssen januari 2007 Vraag 3</u>

Verklaar de noemer in de formule van de oDi

De noemer (k-1)/4 is de waarde voor een minimale spreiding (als alle onderzoekselementen binnen één enkele categorie zitten). Wanneer we deze waarde in de noemer zetten is het

resultaat beter te interpreteren omdat we dan een genormeerde dispersiemaat hebben, waarbij het maximum gelijk is aan 1 en het minimum = 0.

Examen Tijssen januari 2008 Vraag 1

Gegeven: tabel:

	f_i	F_i	F_i^*	Frequentie-dichtheid
1-5	12			
6-10				
11-15	33			
16-20	45	107		
21-25				
Totaal			1,00	24

Gevraagd:
a) Vul de frequenties in de bovenstaande tabel verder aan.

b) Bereken de lineair geïnterpoleerdeinterkwartiele spreidingscoëfficiënt

c) Tussen welke grenzen zal volgens de stelling van Chebychev ten minste 75% van de waarnemingen zich bevinden?

d) Hoeveel procent van de waarnemingen ligt feitelijk tussen deze grenzen?

e) Hoe verklaar je het verschil?

Oplossing:

a) Vermits F_i voor de klasse 16-20 = 107, kunnen we de frequentie van 6-10 berekenen. 107 is nl. de som van de eerste 4 klassen. We vinden dus: 107-45-33-12 = 17. We kunnen nu ook de eerste drie F_i berekenen. Verder weten we dat 24 = frequentiedichtheid, die gedefinieerd wordt als de absolute frequentie/klassebreedte. De klassebreedte is 5. Daaruit kunnen we de absolute frequentie van het totaal berekenen: nl. 5 * 24 = 120. Nu kunnen we al de rest invullen.

	f_i	F_i	F_i^*	Frequentie-dichtheid
1-5	12	**12**	**0,10**	**2,4**
6-10	**17**	**29**	**0,24**	**3,4**
11-15	33	**62**	**0,51**	**6,6**
16-20	45	107	**0,89**	**9**
21-25	**13**	**120**	**1,00**	**2,6**
Totaal	**120**		1,00	24

b) Voor de geïnterpoleerde interkwartiele spreidingscoëfficiënt gebruik je volgende formule en vervang je r door de formulie voor eerste en derde kwartiel en i door Q_1 of Q_3

$$\tilde{x}_r = l_i^e + \frac{r - F_{i-1}}{f_i} . v_i$$

Om dus het eerste geïnterpoleerde kwartiel te berekenen krijg je volgende formule:

$\tilde{Q}_1 = l_{Q_1}^e + \frac{(N+1)/4 - F_{Q_1-1}}{f_{Q_1}} . v$ (v is de klassebreedte van de mediaanklasse)

en het derde kwartiel: $\tilde{Q}_3 = l_{Q_3}^e + \frac{3(N+1)/4 - F_{Q_3-1}}{f_{Q_3}} . v$

Rangorde 1ste kwartiel = (N+1)/4 = 121/4 = 30,25ste waarneming, dit is in de derde klasse

Rangorde 3de kwartiel =3 (N+1)/4 = 90,75ste waarneming, dit is in de vierde klasse

Bepaling van de exacte grenzen. De exacte grenzen zitten net tussenin: dus van 5 als einde van de eeste klasse naar 6 als begin van de volgende klasse: afstand daartussen delen door 2 = 1/2. De exacte bovengrens van de eerste klasse is dus 5,5 en de ondergrens van de volgende is ook 5,5. Om de benedengrens te vinden trek je de afstand die je gevonden hebt (nl. 1/2) af van 1, je bekomt dus als benedengrens 0,5.

	f_i	F_i	F_i^*	Frequentie-dichtheid
0,5-5,5	12	12	0,10	2,4
5,5-10,5	17	**29**	0,24	3,4
10,5-15,5	**33**	**62**	0,51	6,6
15,5-20,5	**45**	107	0,89	9
20,5-25,5	13	120	1,00	2,6
Totaal	120		1,00	24

De gecumuleerde frequentie van de klasse voor de klasse van Q_1 is 29 en de absolute frequentie van de klasse van Q_1 is 33. Voor Q_3 zijn de waarden respectievelijk 62 en 45. De klassebreedte is 5. We hebben nu alle gegevens om de formules voor geïnterpoleerde kwartielen in te vullen:

$\tilde{Q}_1 = 10,5 + \frac{30,25-29}{33} . 5 = 10,68$

$\tilde{Q}_3 = 15,5 + \frac{90,75-62}{45} . 5 = 18,69$

Nu kan je deze waarden gewoon gebruiken in de formule van interkwartiele spreidingscoëfficiënt:

$\frac{\tilde{Q}_3 - \tilde{Q}_1}{\tilde{Q}_3 + \tilde{Q}_1} = \frac{18,69-10,68}{18,69+10,68} = \frac{8,01}{29,37} = 0,27$

c) Voor om het even welke positieve waarde k geldt : minstens een fractie 1 - $-\frac{1}{k^2}$ van alle meetwaarden ligt in het interval]x_{gem}−ks,x_{gem}+ks[

Hiervoor moeten we de standaardafwijking s berekenen en k

Gevraagd is om als fractie 75% te nemen, dus 1 - $-\frac{1}{k^2} = 0{,}75$

Hieruit kunnen we k berekenen. --> - $1/k^2$ = 0,75 -1

--> $1/k^2$ = 0,25

--> k^2 = 1/0,25

--> k^2 = 4 dus k = 2

Dus minstens 75% van de waarnemingen ligt tussen x_{gem}-2.s en x_{gem} +2.s

We berekenen nu het gemiddelde en de standaardafwijking.

Voor het gemiddelde maken we de som van x_i*f_i en voor x_i gebruiken we telkens het klassemidden. Deze som is 1710. Deze delen we door N om het gemiddelde te bekomen: 1710/120 = 14,25.

Om de standaardafwijking te bekomen maken we eerst de verschillen ten opzichte van het gemiddelde. Deze kwadrateren we en vervolgens vermenigvuldigen we deze kwadraten met de frequenties. Ten slotte maken we de som: deze is dan gelijk aan 3982.5. Om de standaardafwijking te bekomen delen we deze som door N-1 en vervolgens nemen we de vierkantswortel: $\sqrt{\frac{3862.5}{119}} = 5{,}697$

	f_i	Klassemidden (x_i)	x_i*f_i	x_i-x_{gem} (verschil tov gemiddelden)	$(x_i$-$x_{gem})^2$ (Kwadraten van verschil)	$f_i(x_i$-$x_{gem})^2$
1-5	12	3	36	-11,25	126,5625	1518,75
6-10	17	8	136	-6,25	39,0625	664,0625
11-15	33	13	429	-1,25	1,5625	51,5625
16-20	45	18	810	3,75	14,0625	632,8125
21-25	13	23	299	8,75	76,5625	995,3125
Totaal	120		1710			3862,5

=$\sum f_i(x_i$-$x_{gem})^2$

Toepassing van de formule Van Chebychev:

75% ligt tussen x_{gem}-2.s en x_{gem} +2.s

Dus 75% ligt tussen 14.25 - 2*5,70 en 14.25 + 2*5,70

75% ligt tussen 2.85 en 25.65

d) Hoeveel procent ligt daadwerkelijk tussen deze grenzen?

Als de frequenties uit de eerste klasse egaal verdeeld zijn, zouden er van de 12 ongeveer de helft (dus 6) lager dan 2,85 liggen, dus buiten de grenzen. Dat betekent dat de rest, nl. 114 metingen of 95% binnen de grenzen liggen.

e) Hoe verklaar je het verschil?

Chebychev geeft enkel aan hoeveel er 'minstens' binnen die klasse liggen, maar het kunnen er meer zijn.

<u>Examen Thijssen januari 2008 Vraag 3</u>

Gegeven: lineaire transformatie van de vorm Y = 2X + 20

standaardafwijking na transformatie Y = 10

gemiddelde na transformatie Y = 100.

Gevraagd: Variatiecoëfficiënt van de oorspronkelijke variabele X?

Oplossing:

Variatiecoëfficiënt (VC) = $\frac{s}{x_{gem}}$

Om het gemiddelde van de oorspronkelijke variabele te vinden moeten we de transformatie maken: 100 = 2x + 20. --> x = 40

Op de standaardafwijking heeft enkel de factor waarmee x werd vermenigvuldigd invloed, dus 10 = 2x --> x = 5

VC = 5/40 = 0.125

➔ Antwoord A

<u>Examen Thijssen januari 2012 Vraag 1</u>

Gegeven:

	f_i	F_i	F_i^*	Frequentie-dichtheid	Klassencentra
1-10				1,5	
11-20	50				
21-30			0.80		
31-40					
Totaal	100				

Gevraagd:

a) Bepaal het verschil tussen het lineair geïnterpoleerde negentigste en tiende percentiel.

b) Wat leer je uit voorgaande uitkomst?

c) Bereken op basis van de stelling van Chebychev de grenzen waartussen zich tenminste 70% van de waarnemingen bevinden.

d) Teken het ogief en bepaal grafisch de mediaan

Oplossing: Vul de tabel verder aan: begin met de berekening van f_i dmv de frequentie-dichtheid: 1,5 * 10 = 15 .

	f_i	F_i	F_i^*	Frequentie-dichtheid	Klassencentra
1-10	**15**	**15**	**0,15**	1,5	**5,5**
11-20	50	**65**	**0,50**	**5**	**15,5**
21-30	**15**	**80**	0.80	**1.5**	**25,5**
31-40	**20**	**100**	1	**2**	**35,5**
Totaal	100				

a) Voor verschil tussen het lineair geïnterpoleerde negentigste en tiende percentiel gebruik je volgende formule:

$$\tilde{x}_r = l_i^e + \frac{r - F_{i-1}}{f_i} . v_i$$

en je vervangt r door de formule voor percentiel en i door P_{90} en P_{10}. De formule voor 90ste percentiel is P_{90} = 90(N+1)/100 en voor 10de = P_{10} = 10(N+1)/100. Uit deze formule weten we dat het 90ste percentiel zich bevindt op de 90,90 ste plaats en het 10de op de 10, 1de plaats. Dat is in de vierde en eerste klasse.

$\tilde{P}_{90} = l_i^e + \frac{90(N+1)/100 - F_{P_{90}-1}}{f_{P_{90}}} . v$ (v is klassebreedte van 90ste percentiel)

en

$\tilde{P}_{10} = l_i^e + \frac{10(N+1)/100 - F_{P_{10}-1}}{f_{P_{10}}} . v$ (v is klassebreedte van 10de percentiel)

Bepaal de exacte grenzen: deze zitten exact tussen het einde van klasse 1 en het begin van klasse 2. Vermits er 1 tussen zit, is de helft ervan 0,5. De exacte bovenklasse van de eerste klasse is dus 10+0,5 = 10,5. Om de benedengrens te vinden trekken we een half af van 1. De exacte ondergrens van de eerste klasse is dus 1-0,5 = 0,5. Bij de vierde klasse is de exacte ondergrens 30,5. Vervolgens bepalen we de gecumuleerde frequentie van de voorgaande klasse: bij de eerste klasse is dat 0 en bij de laatste klasse is dat 80. Verder hebben we de absolute frequenties nodig van klasse 1 en 4: dat is respectievelijk 15 en 20. De klassenbreedte is 10. We kunnen nu alle waarden in de formules invullen:

$\tilde{P}_{90} = 30{,}5 + \frac{90{,}90 - 80}{20} . 10 = 35{,}95$

$\tilde{P}_{10} = 0{,}5 + \frac{10{,}1 - 0}{15} .10 = 7{,}23$

Het verschil is dus 28,72

b) Wat leer je uit voorgaande uitkomst?

80% van de gevallen zit tussen 7,23 en 35,95.

c) Bereken op basis van de stelling van Chebychev de grenzen waartussen zich tenminste 70% van de waarnemingen bevinden.

Voor om het even welke positieve waarde k geldt : minstens een fractie 1 - $\frac{1}{k^2}$ van alle meetwaarden ligt in het interval $]x_{gem}-ks, x_{gem}+ks[$

Hiervoor moeten we de standaardafwijking s berekenen en k

Gevraagd is om als fractie 75% te nemen, dus 1 - $\frac{1}{k^2} = 0{,}70$

Hieruit kunnen we k berekenen. --> $- 1/k^2 = 0{,}70 - 1$

--> $1/k^2 = 0{,}30$

--> $k^2 = 1/0{,}30$

--> $k^2 = 3{,}333$ dus $k = 1{,}8$

Dus minstens 70% van de waarnemingen ligt tussen x_{gem}-1,8.s en x_{gem} +1,8.s

We berekenen nu het gemiddelde en de standaardafwijking.

Voor het gemiddelde maken we de som van $x_i * f_i$ en voor x_i gebruiken we telkens het klassemidden. Deze som is 1950. Deze delen we door N om het gemiddelde te bekomen: 1950/100 = 19,50.

Om de standaardafwijking te bekomen maken we eerst de verschillen ten opzichte van het gemiddelde. Deze kwadrateren we en vervolgens vermenigvuldigen we deze kwadraten met de frequenties. Ten slotte maken we de som: deze is dan gelijk aan 9400. Om de standaardafwijking te bekomen delen we deze som door N-1 en vervolgens nemen we de vierkantswortel: $\sqrt{\frac{9400}{99}} = 9{,}7$

	f_i	Klassemidden (x_i)	$x_i * f_i$	$x_i - x_{gem}$ (verschil tov gemiddelden)	$(x_i - x_{gem})^2$ (Kwadraten van verschil)	$f_i(x_i - x_{gem})^2$
1-10	15	5,5	82,5	-14	196	2940
11-20	50	15,5	775	-4	16	800
21-30	15	25,5	382,5	6	36	540
31-40	20	35,5	710	16	256	5120
Totaal	100		1950			9400

$=\sum f_i(x_i - x_{gem})^2$

Toepassing van de formule Van Chebychev:

70% ligt tussen x_{gem}-1,8.s en x_{gem} +1,8.s

Dus 70% ligt tussen 19,5 - 1,8*9,7 en 19,5 + 1,8 * 9,7

70% ligt tussen 2.04 en 36,96

d) Teken het ogief en bepaal grafisch de mediaan

Om het ogief te tekenen gebruik je gecumuleerde frequenties voor de y-as en exacte klassegrenzen voor de x-as. Om de ogief van nul te laten starten gebruik je als eerste exacte grens de ondergrens en voor de rest de exacte bovengrenzen. De mediaan vind je door de waarde (N+1)/2 = 50,5 af te meten op de y-as en een lijn te trekken tot de grafiek. Op de x-as vind je de waarde van de mediaan.