Korte Samenvatting van H7.

(Versie 1.0 Bart van Zweeden, MLA: 2024-2025)

VK Voorkennis en begrippen is al een beetje een herhaling van H2.

Centrummaten:

- Gemiddelde. Een bekende.
- Mediaan De middelste van een rij op volgorde gezette getallen.
- Modus Het meest voorkomende getal. Soms is ie er niet.

Spreidingsmaten:

- Spreidingsbreedte Grootste – Kleinste waarneming/getal.
- **Standaardafwijking** De gemiddelde afwijking van de getallen ten opzichte van het berekende gemiddelde.
  Deze berekening (als ie al berekend moet worden) is typisch iets voor de GRM.

En deze:

De populatieproportie p van een kenmerk in de populatie is

$$p = \frac{\text{aantal elementen met het kenmerk in de populatie}}{\text{totaal aantal elementen in de populatie}}.$$

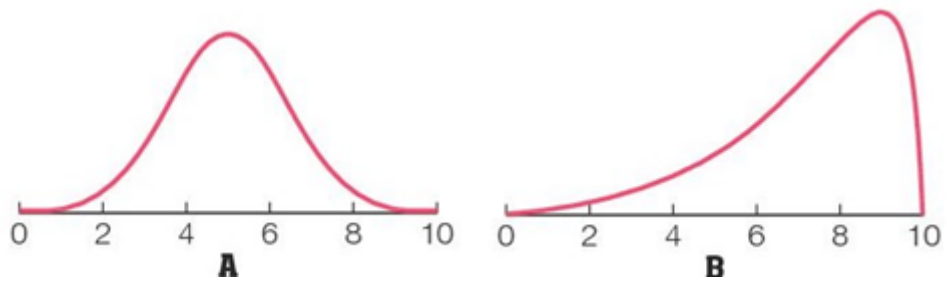
De steekproefproportie $\hat{p}$ van een kenmerk in de steekproef is

$$\hat{p} = \frac{\text{aantal elementen met het kenmerk in de steekproef}}{\text{totaal aantal elementen in de steekproef}}.$$

7.1 Statistische verdelingen.

De paragraaf opent met een klein overicht van verdelingen. Eigenlijk blijkt later dat er maar drie echt van belang zijn:

1. Normaal verdeeld. (plaatje A)
2. Links scheef verdeeld. (plaatje B.)
3. Rechtsscheef verdeeld. (laat zich raden: het spiegelbeeld van B).



Wat is een normale verdeling? Hoezo normaal?

Normaal verdeeld wijst naar variabelen die merendeels in de natuur (wij ook…) voorkomen en naar variabelen van objecten waar wij als mensen mee te maken hebben, door ons gemaakt zijn.
De eerste categorie: Alle lichaamslengten, alle lichaamsgewichten, omtrekken van menselijke schedels, lengten opperarm beentje bij kinderen, gewichten van fruit uit boomgaarden, gewichten kippeeieren, etc etc etc.
Ook (tweede categorie): het vulgewicht van pakken hagelslag, inhoud van blikken soep, de diameter van geproduceerde stalen zuigers voor in een motorblok, het gehalt ethanol in benzine bij verschillende merken etc etc etc….

Elke normale verdeling kan je beschrijven met slechts twee grootheden:

1. **Het gemiddelde**
2. **De bijbehoren standaarddeviatie (standaardafwijking).**

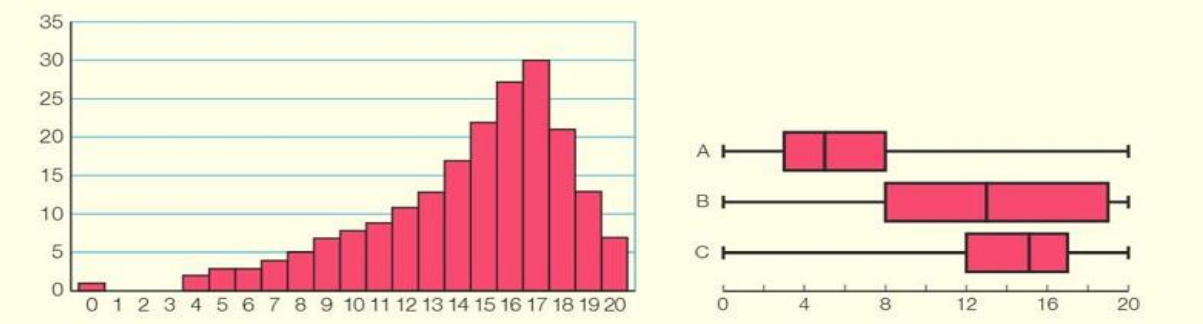
Eis is wel dat eerder onderzocht is dat de genoemde variabele ook echt normaal verdeeld is.

SITUATIE:
Een onderzoeker meet het gehalte van een chemische verbinding voorkomend in pindakaas. Het gehalte bij Calvé is gemiddeld 17gram bij een standaardafwijking van 0,9 gram.
Bij AH-eigen merk is dat 16,1 gram bij een standaardafwijking van 1,3 gram.
Je zou (…) met de effectgrootte kunnen kijken of dit verschil klein, middelmatig of groot is. Daar gaat het nu even niet om. (Mag je zelf als oefening wel ff checken…)
Straks zien we wat het vervolg hiervan is.

Het boek neemt de afslag om de verdelingen (feitelijk niks anders dan histogrammen die het resultaat zijn van klasseindelingen) te verbinden met boxplots.

Links en Rechts scheef verdeeld.

Betekend dat de grafiek een uitslag naar links heeft.
Rechts scheef verdeeld: een uitslag naar rechts.



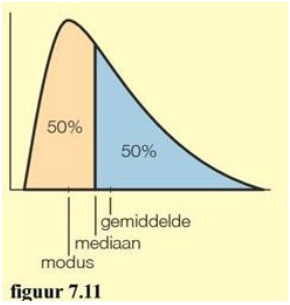
Linker plaatje: de links-scheve verdeling. Welke boxplot hoort erbij en waarom?

A. 75% van de data zit onder de 8. Dat is bij de verdeling NIET het geval. Nee dus
B. 75% van de data zit voorbij de 8. Zou B kunnen zijn, maar het rode blok met 50% van de data is te breed. Tussen de 8 en bijna 19. **Dus toch plaatje C.**

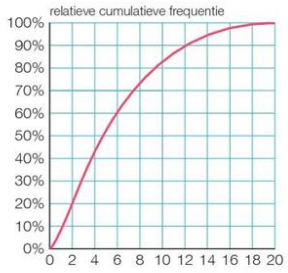
Het is een beetje inschatten. Da's duidelijk. Op een examen moet je dus echt ook letten op de getallen en niet alleen afgaan op je gevoel..

Tenslotte: Meertoppige verdelingen kunnen worden veroorzaakt doordat twee onderzoeksgroepen (per ongeluk?) bij elkaar geveegd zijn. Meisjes-jongens, of de appels van twee boomgaarden.. Dat maakt uitspraken doen lastig en dus onwenselijk.
Soms weet je als onderzoeker niet beter en ontdek je dat pas bij de analyse.
Oefen met opdr 4 en 5, pag 92,93.
Tenslotte:

- Bij de normale verdeling liggen mediaan en gemiddelde op dezelfde plek.
- Bij de link scheve verdeling ligt het gemiddelde links van de mediaan en links van de modus (is het hoogste punt) .Merk op: een mediaan hakt de verzameling gegevens precies in tweeën! Diens plek is niet zomaar te zien in een plaatje. Wel kan je zeggen dat oppervlakte links = oppervlakte rechts. Het gemiddelde wordt beïnvloed door de uitschieters. Denk aan de miljonair tussen 100 mensen met een laag inkomen. Hij trekt in z'n eentje het gemiddelde inkomen erg omhoog..
- Bij de rechtsscheve verdeling ligt het gemiddelde rechts van de mediaan.



figuur 7.11



figuur 7.12

Bij dit plaatje (zie boek) is de vraag welke verdeling erachter zit. Mijn truc: kijk bij 50%. Daar zit de mediaan: Juist bij waarde 5 dus.
De modus is dat stuk met het steilste deel. Dat is slecht te zien. Wel kan je snappen dat na de waarde vijf nog vele getallen volgen: dus rechtsscheef.

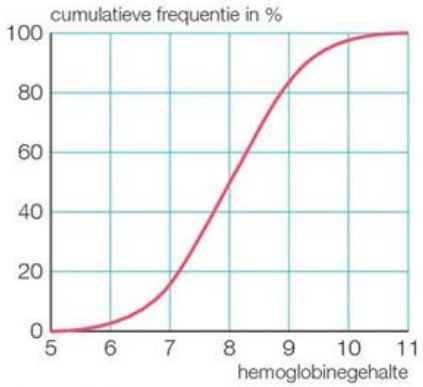
10

Bij een medisch onderzoek is het hemoglobinegehalte van een grote groep Nederlandse mannen gemeten. Zie de relatieve cumulatieve verdelingskromme in figuur 7.15.

a Met welke soort verdeling heb je te maken? Licht toe.

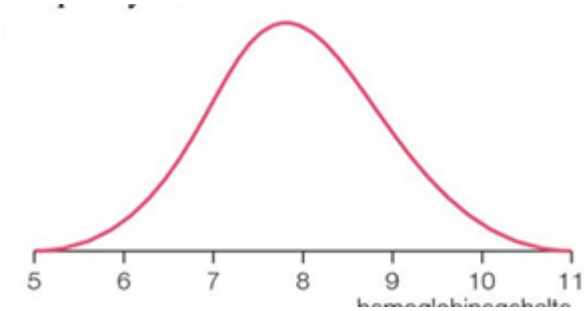
b Schets de bijbehorende verdelingskromme. Neem op de horizontale as de getallen 5, 6, 7, 8, 9, 10 en 11.

c Schat het gemiddeld hemoglobinegehalte van de onderzochte groep.



figuur 7.15

Uit het boek een goede vraag. Het plaatje toont de cumulatieve frequentie die hoort bij de symmetrische normale verdeling. De modus, mediaan en gemiddelde liggen op hetzelfde punt: bij 50% en dus de waarde 8. Gemiddeld 8 dus, modus: meest voorkomende klasse is rondom diezelfde 8 en de mediaan ligt altijd bij 50%.

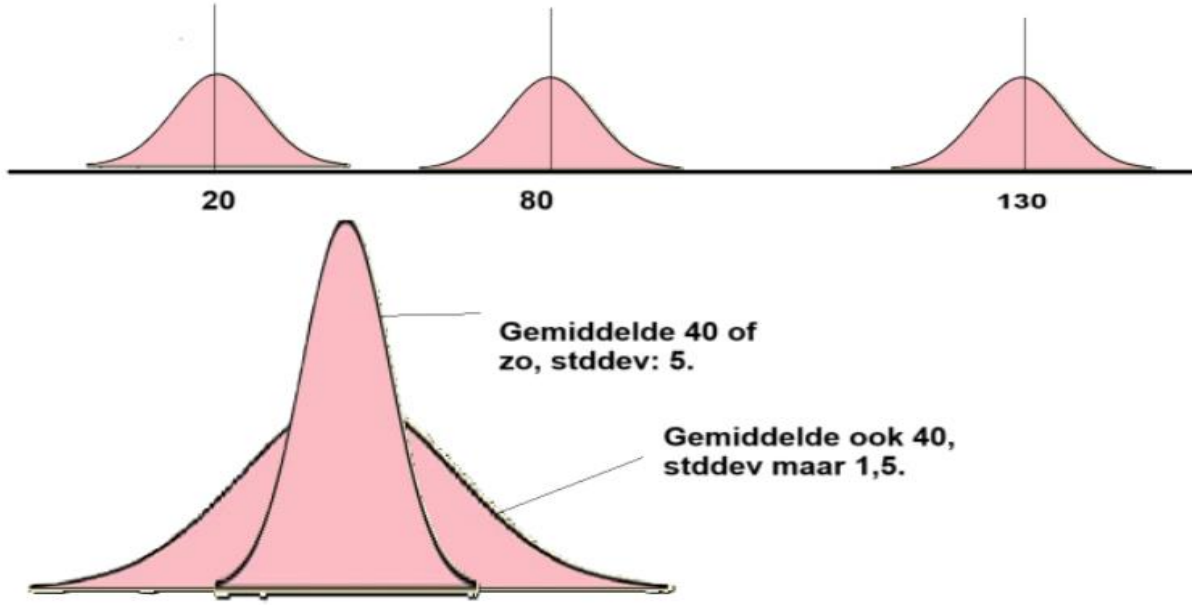


Uit de uitwerkingenbundel.
Dit is het kenmerkende beeld van de normale verdeling.

En dan nu:

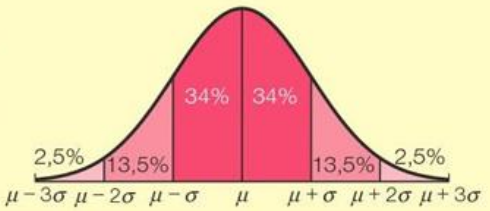
DE NORMALE VERDELING

Zoals gezegd: een klokcurve zoals dat heet. Bell-shaped zeggen de Engelsen.
Deze curve die dus de uitkomstgrafiek is van veel statistiek is ontwikkeld door Carl Friedrich Gauss. Deze meer dan geniale man heeft een wiskundige formule (en grafiek) weten te ontwikkelen die alle normaal verdeelde curves van elkaar verschillen en met elkaar gemeen hebben.
De **plaats** van het gemiddelde: Die kan op verschillende plekken liggen, bijv gemiddeld 20,80 of 130 gram… Aangezien de curves hetzelfde zijn, is hun standaardafwijking wel steeds dezelfde. Mss 7 of zo. (gokje). Ook de **breedte** van de curve is bepalend. Hoe breder, hoe meer variatie.



In dit onderste plaatje is het gemiddelde van beide partijen op 40. Ze verschillen in standaardafwijking. Voorbeeld: de smalle curve is de populatie appels bij een dure groenteboer: alle appels wegen gemiddeld 40 gram maar hun onderlinge verschillen zijn gering. De brede curve is lager en breder: grotere variatie in de partij die mss net van het land komt.
Onderstaande afbeelding is HEEEEEEL belangrijk:

Vuistregels bij de normale verdeling
- ongeveer 68% van de waarnemingsgetallen ligt tussen $\mu - \sigma$ en $\mu + \sigma$
- ongeveer 95% van de waarnemingsgetallen ligt tussen $\mu - 2\sigma$ en $\mu + 2\sigma$
- bijna 100% van de waarnemingsgetallen ligt tussen $\mu - 3\sigma$ en $\mu + 3\sigma$



figuur 7.24

Elk vak is precies de afmeting van de standaard afwijking. Dat kenmerkt de normale verdeling.
De middelde rode vakken zijn samen ca 68% van de waarnemingen/metingen. Dan volgen twee vakken van elk 13,5% en tenslotte de laatste vakjes elk 2,5%.

Om die grenzen gaat het vaak: 95% van alle metingen vallen binnen het gemiddelde + en – twee maal de standaardafwijking.

7.2 Betrouwbaarheidsintervallen.

Een lastig deel van het hoofdstuk.
Wat is een betrouwbaarheidsinterval? Het woord zegt het al: het interval van getallen waarbinnen zich 68 of 95% van de waarnemingen of metingen bevinden en die iets zeggen over hoé betrouwbaar je onderzoek is.
Een voorbeeld:
Een arts onderzoekt 70 kinderen van ca 10 jaar op hun gewicht en vindt een gemiddelde van 35kg met een standaardafwijking. Van 5 kg. Landelijk liggen deze cijfers op 39,5kg met een standaardafwijking van 7,1 kg. Is zijn onderzoek betrouwbaar? (Juist ja, de effectgrootte zou je kunnen inzetten).
Maar nu dit: Hij neemt in een ander dorp ook een steekproef van ca 10 kinderen en vindt nu een gemiddelde van 55kg. Standaardafwijking is 8 kg. Betrouwbaar?
Vermoedelijk is de steekproef te klein. Hij heeft te maken met een fikse afwijking.

Steekproeven moeten vaststellen wat een landelijk cijfer is: je kan onmogelijk alle kinderen van 10 jaar gaan wegen. Merk op dat bij consultatiebureaus dit bij baby's dus wél gebeurd, maar dat terzijde.

Men doet dus steekproeven.
Wat kenmerkt een steekproef? Een gemiddelde en een standaardafwijking.
Hoe doet men dit?

- Het gemiddelde wordt op een simpele manier vastgesteld.
- De standaardafwijking die bij deze steekproef hoort, wordt berekend op de manier die je al kent. Pak de GRM dus. Echter bij een steekproef is de standaardafwijking kleiner dan die van de gehele populatie.

> **Bij een steekproef van n objecten uit een normale verdeling met gemiddelde μ en standaardafwijking σ is het steekproefgemiddelde normaal verdeeld met gemiddelde μ en standaardafwijking $\frac{\sigma}{\sqrt{n}}$.**

Die laatste wordt aangeduid met een hoofdletter S.
Wie een steekproef neemt uit een groter geheel, zal een gemiddelde vinden dat in de buurt ligt van het echte gemiddelde. Dat is niet zo gek.
De standaardafwijking die bij deze steekproef hoort is afhankelijk van hoeveel metingen we deden: S= $\frac{\sigma}{\sqrt{n}}$. Oorzaak: veel voorkomende elementen in de populatie hebben een grotere kans om in de steekproef terecht te komen waardoor hun onderlinge verschillen kleiner zijn.

Nog zo'n opmerking: Als je meermalen een steekproef uitvoert krijg je dus ook steeds een steekproefgemiddelde die net effe anders is dan de vorige. **Al die steekproefgemiddelden zijn zelf ook weer normaal verdeeld!**

VOORBEELD:
In de supermarkt vinden we bakjes tomaatjes. Elke bakje weegt 400 gram bij een standaardafwijking van 21 gram. Hoe is de verdeling als je een steekproef neemrt van 9 van die bakjes tomaatjes?

1. We gaan er vanuit dat het gemiddelde gewicht van de bakjes normaal verdeeld is met een gemiddelde van 400 gram, zo ook dus de steekproef.
2. De steekproefstandaardafwijking is nu S=21/V9 = 21/3 = 7 gram.
3. Dat betekent dat 68% van de bakjes een gewicht zal hebben tussen de 400 – 7 tot 400 + 7 ofwel tussen de 393 en 407 gram.
   95% van de bakjes zitten tussen de 386 en 414 gram gemiddeld.

Oefen met sommen van ca 20 tot ca 25.

Waar draait het nu om?

Je weet in de praktijd niet wat de maten zijn van de gehele populatie.
Dus door het doen van steekproeven kom je snel te weten wat het steekproefgemiddelde is, en het daarbij berekende standaardafwijking.
We zeggen nu dit: Er is 95% kans dat de echte waarde van het populatiegemiddelde van de hele populatie zal liggen tussen de grenzen:

Het 95%-betrouwbaarheidsinterval voor het populatiegemiddelde is $\left[\overline{X} - 2 \cdot \frac{S}{\sqrt{n}}, \overline{X} + 2 \cdot \frac{S}{\sqrt{n}}\right]$. Hierin is $\overline{X}$ het steekproefgemiddelde, S de steekproefstandaardafwijking en n de steekproefomvang.

X met streepje is het steekproefgemiddelde. S is de steekproefstandaardafwijking.

Een duidelijk effect is de omvang van de steekproef.

Iemand meet een gemiddelde van 300 bij een steekproefstd-afw van 25. Omvang 100.
Dan volgt eruit dat de 95% grenzen waarbinnen de echte u zit, zijn:
300 – 2x 25/V100 = 300 – 5. Deze grenzen zijn dan 295 en 305. Je weet dan voor 95% zeker dat het echte gemiddelde van de hele populatie tussen de 295 en 305 ligt. Ja er is dus altijd een kleine kans van 5% dat dat niet zo is.
Iemand die niet 100 maar 1000 objecten onderzoekt, kan hetzelfde gemiddeld van 300 vinden. Zijn grenzen zijn echter: 300 – 2 x 25/V1000 = 298,4 en 301,6. Je ziet dat het interval krapper wordt: door meer objecten, personen, dingen etc worden de uitspraken betrouwbaarder: Het 95% interval is nu krapper ronde de 300: we weten zekerder dat het echte gemiddelde dicht bij de 300 ligt. Test jezelf even met opdracht 26 en 29.

26 Bereken het 95%-betrouwbaarheidsinterval voor het populatiegemiddelde bij

a $\overline{X} = 100$; $S = 24$ en $n = 144$

b $\overline{X} = 80{,}5$; $S = 15{,}5$ en $n = 100$

c $\overline{X} = 12{,}50$ en $\frac{S}{\sqrt{n}} = 0{,}65$.

7.3 Betrouwbaarheidsintervallen bij proporties.

In de herhaling:
Twee belangrijke begrippen:

Populatieproportie

Het gedeelte van de populatie dat voldoet aan een kenmerk. Vb: het percentage Nederlanders die blauwe ogen hebben. Dit kan dus zijn: p=0,763
Dat betekent dat 76,3% van de Nederlands blauwe ogen heeft.

Steekproefproportie

De betekenis is dezelfde, echter kijkt men nu naar een kleiner gedeelte: Stel je doet een steekproef op school (1200 leerlingen) en je ontdekt dat daar een p=0,698 uit komt. Dat betekent dat op school er 69,8% van de leerlingen blauwe ogen hebben.
Waarom zijn die getallen niet gelijk? Laten we duidelijk zijn: alles wat we weten over enorm grote populaties zoals de bevolking van een land IS gebaseerd op steekproeven!!
Wat is de clue: Door meerdere steekproeven te nemen, krijg je steeds verschillende waarden voor het steekproefgemiddelde. Die uitkomsten zitten vaak best dicht bij elkaar én dicht bij de waarde die als 'landelijk' kan worden aangenomen. Hierover later meer.

Eigenlijk gaat het in deze situatie om hetzelfde: je wil weten wat de proportie is van hele populatie, op basis van steekproeven. In plaats van gemiddelden hebben we het nu over proporties.

Landelijk is bekend dat proportie brildragende automobilisten ca 19% is. (Volkomen bedacht hoor..)
Ik doe nu een steekproef van 100 automobilisten en ga kijken hoeveel ervan er een bril dragen. Ik vind 23 mensen. De populatieproportie is dus 0,19 en de steekproefproportie is nu 0,23. Balen: dat zou hetzelfde moeten zijn. Maar tussen welke grenzen mag ik zitten zodat deze uitkomst toch niet afwijkend is?

We stellen de steekproefstandaardafwijking als volgt vast:

Bij populatieproportie p en steekproefomvang n is de steekproefproportie normaal verdeeld met gemiddelde p en standaardafwijking $\sigma = \sqrt{\frac{p(1-p)}{n}}$.

We vonden dus een p=0,23. n=100 en dus sigma (omgevallen zesje) =V(0,23(1-0,23)/100) = 0,04. Op basis van deze steekproef neem ik aan dat de populatieproportie ligt tussen 2 x 0,04 links en rechts van de 0,23: tussen grenzen: 0,15 en 0,31.

Getal en Ruimte draait het een beetje om en dat kan verwarrend werken.

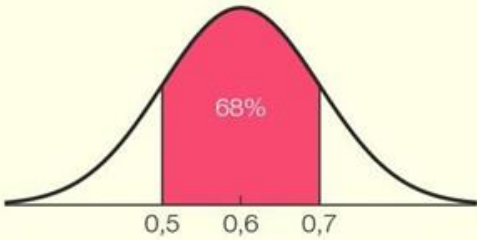
Voorbeeld

In een land heeft 60% van de bevolking bruine ogen. Bij een groep van 24 personen wordt de proportie personen met bruine ogen genoteerd. Bereken bij welk percentage van de mogelijke steekproeven de steekproefproportie tussen 0,5 en 0,7 zal liggen.

Uitwerking

$\mu = p = 0{,}6$ en $\sigma = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0{,}6 \cdot 0{,}4}{24}} = 0{,}1.$

Zie de schets, dus bij 68%.



Je kan hier dus zeggen dat bij 68% van de steekproeven van 24 mensen (meerdere steekproeven dus), de uitkomsten liggen tussen de 0,5 en 0,7.

In de pagina (114) leggen ze de zaak beter uit.

1. Je doet meerdere steekproeven, die uit n personen bestaan
2. Elke steekproef heeft zijn eigen steekproefgemiddelde en standaardafwijking die je met die formule vaststelt.
3. Aangezien de steekproefgemiddelden zelf ook normaal verdeeld zijn (!!!) geldt dat 95% van die uitkomsten zit tussen de grenzen p-2s en p+2s.
4. Andersom: je hebt 5% kans dat er een steekproefresulaat uitkomt dat teveel afwijkt van het landelijke proportie.

Het voorbeeld op pag 115 is wel helder:

Voorbeeld

Twee dagen voor een verkiezing wordt aan 1200 stemgerechtigde Nederlanders gevraagd of ze van plan zijn te gaan stemmen. Van deze 1200 zeggen 756 inderdaad te gaan stemmen. Bereken het 95%-betrouwbaarheidsinterval voor de proportie stemgerechtigde Nederlanders die van plan zijn te gaan stemmen.

Uitwerking

$\hat{p} = \frac{756}{1200} = 0{,}63$ en $\sigma = \sqrt{\frac{0{,}63 \cdot 0{,}37}{1200}} = 0{,}0139...$

$\hat{p} - 2\sigma = 0{,}63 - 2 \cdot 0{,}0139... \approx 0{,}602$ en

$\hat{p} + 2\sigma = 0{,}63 + 2 \cdot 0{,}0139... \approx 0{,}658$

Het 95%-betrouwbaarheidsinterval is [0,602; 0,658].

Rond in een betrouwbaarheidsinterval proporties af op drie decimalen, tenzij anders wordt gevraagd.

Wat zegt deze uitkomst? Er is een kans van 95% dat er tussen de 602 en 658 mensen van de 1200 gaan stemmen. Moet je meer weten? Dan moet je meer mensen ondervragen.

Een actuele:
Een exitpoll onder 10000 Amerikanen liet zien dat het aantal Harris-stemmers lag met een steekproefproportie van 0,48. Zou zijn op grond van deze gegevens de verkiezingen

gewonnen hebben?

1. P = 0,48, de steekproefstandaardafwijking is 0,00499.
2. Het betrouwbaarheidsinterval ligt tussen 0,48 – 0,00999 en 0,48 + 0,00999 = [0,47m 0,49] afgerond. Met deze uitslag zou zij net niet gewonnen hebben.. We kennen de uitslag inmiddels...

Oefen met opgaven 38,39,40 om het goed te begrijpen.

7.4 Betrouwbaarheisintervallen toepassen.

Een klein zinnetje eruit gelicht:

Voor het betrouwbaarheidsinterval voor een populatieproportie heb je $\hat{p}$ en n nodig.
Voor het betrouwbaarheidsinterval voor het populatiegemiddelde heb je $\overline{X}$, S en n nodig.

Er zijn er dus twee. Ja. Maar die verschillen niet wezenlijk. Wat belangrijk is, is dat je in examensituaties door hebt of men spreekt over proporties, soms uitgedrukt als percentage, of over gemiddelden.
Het verschil:

- 64% van de mannen in Nederland heeft kenmerk X. p = 0,64 dus.
- De gemiddelde Nederlandse man weeg gemiddeld 81 kg.

Bij de eerste bereken je sigma met die grote wortel, bij de tweede gebruik je S/Vn.
In beide gevallen is n, de steekproefomvang betrokken. Ja logisch.
Dit staat op je examenblad en op het SE.

Op het formuleblad bij het centraal examen staan de formules voor de 95%-betrouwbaarheidsintervallen als volgt genoteerd.

Betrouwbaarheidsintervallen

Het 95%-betrouwbaarheidsinterval voor de populatieproportie is $p \pm 2 \cdot \sqrt{\frac{p(1-p)}{n}}$, met p de steekproefproportie en n de steekproefomvang.

Het 95%-betrouwbaarheidsinterval voor het populatiegemiddelde is $\overline{X} \pm 2 \cdot \frac{S}{\sqrt{n}}$, met $\overline{X}$ het steekproefgemiddelde, n de steekproefomvang en S de steekproefstandaardafwijking.

Leer ermee om te gaan.

Nog zo eentje:

Er zijn drie opmerkelijke verschillen tussen de notatie op het formuleblad bij het centraal examen en de notatie in dit boek.

1 In plaats van intervalnotatie wordt een notatie met ± gebruikt.
2 Voor de steekproefproportie wordt p in plaats van $\hat{p}$ gebruikt.
3 Op de plek van σ staat $\sqrt{\frac{p(1-p)}{n}}$.

Opgaven 50 tot en met 58 is echt zinvol om nog een keer na te lopen.

Bij de laatste theorie draait men de zaak om: Hoe groot moet de steekproefomvang zijn, om een bepaald interval te krijgen? Hoeveel mensen moet je ondervragen om het antwoord binnen zekere grenzen te hebben?

Na een onderzoek van n personen ontstaat een proportie van 0,7. Het 95% betrouwbaarheidsinterval ligt tussen 0,65 en 0,75.
Hoe groot is n geweest?
Eigenlijk een idiote vraag. Je wéét toch hoeveel mensen geïnterviewd zijn? Nou ja.
Aanpak:

1. Elke 95% interval is vier-sigma (vier steekproefstandaardafwijkingen) breed.
2. Dus tussen 0,65 en 0,75 ligt 0,1 afstand. Dat maakt dat sigma was: 0,025.
3. Nu de formule:
   0,25 = V(0,70 (1-0,7)/n

   Truc die altijd en snel werkt:
   $0{,}25^2 = 0{,}70(1-0{,}70)/n$
   $n = 0{,}70*0{,}30 / 0{,}25^2 = 336$. Geen gedoe met de GR dus.

Lees ook mijn ‘alternatieve’ samenvatting specifiek van dit verhaal.
Je hebt er nu dus 2.