Statistiek voor MBW / KW
Uitwerkingen huiswerkopgaven

Dr. ir. D.A.M.P. Blom & Dr. J.B.M. Melissen

2025

Week 11: chikwadraatverdeling

Hoofdstuk 10

Opdracht 10.m1: Voor een kansvariabele die een theoretische chikwadraatsverdeling met vijf vrijheidsgraden volgt, geldt dat ...

(a) deze symmetrisch rondom 0 ligt.

(b) deze symmetrisch rondom 5 ligt.

(c) de kansen hiervoor kunnen worden gevonden als het kwadraat wordt genomen van een normaal verdeelde variabele met $\mu = 5$.

(d) deze kansvariabele waarden kan aannemen die groter zijn dan 15.

Uitwerking

Het juiste antwoord is (d).

Opdracht 10.m2: Bij een chikwadraattoets voor onafhankelijkheid moet worden gewerkt met de chikwadraatverdeling met 6 vrijheidsgraden. Er moet worden getoetst met $\alpha = 0,05$. De kritieke tabelwaarde is daarom ...

(a) $1,64$

(b) $14,45$

(c) $12,59$

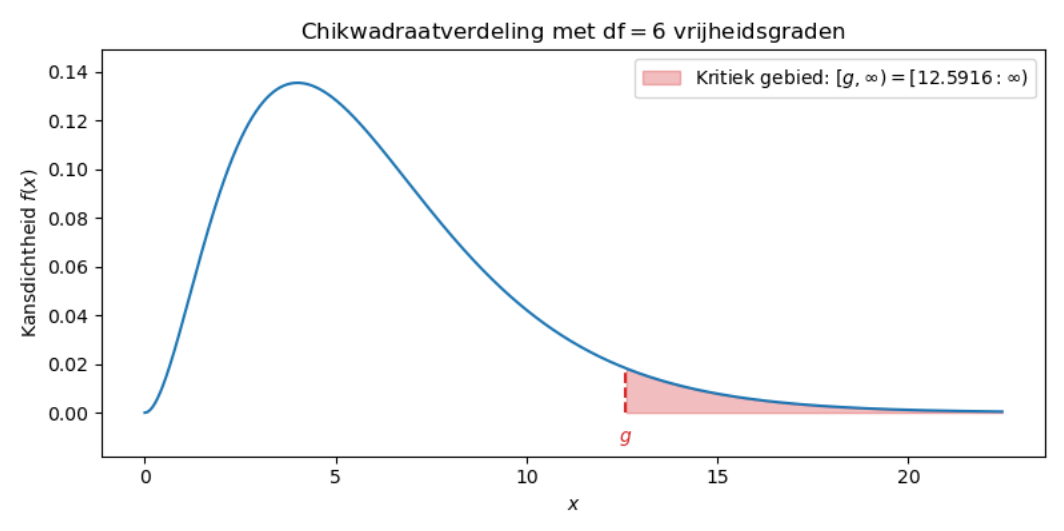
(d) $9,49$

Uitwerking

Laat X de toetsingsgrootheid zijn die chikwadraat verdeeld is met df$= 6$ vrijheidsgraden. In dit geval is de kritieke waarde de waarde x waarvoor geldt dat $P(X \geq x) = 0,05$. Deze kritieke waarde berekenen we met de solver optie als volgt:

$$y_1 = \chi^2\text{cdf}(\text{lower} = x; \text{upper} = 10^{10}; \text{df} = 6)$$
$$y_2 = 0,05$$

De numerieke solver geeft (afgerond op vier decimalen) $x \approx 12,5916$. Het juiste antwoord is dus (c).



Opdracht 10.m3: Bij een toets voor onafhankelijkheid geeft de tabel de waargenomen frequenties weer. De te verwachten of *expected* frequentie voor de cel waarin 80 waarnemingen vermeld zijn, bedraagt . . .

(a) $\frac{80}{250}$

(b) 125

(c) $0,2$

(d) 100

	Man	**Vrouw**
Links	80	120
Rechts	170	130

Uitwerking

Bekijk nu een uitgebreidere versie van de tabel van *observed* frequenties hierboven, waarbij we ook de rij- en kolomtotalen hebben toegevoegd.

	Man	**Vrouw**	**Totaal**
Links	80	120	200
Rechts	170	130	300
Totaal	250	250	500

De te verwachten of *expected* frequentie E_{ij} voor de cel in rij i en kolom j is gelijk aan

$$E_{ij} = \frac{\text{rijtotaal}_i \cdot \text{kolomtotaal}_j}{\text{totaal}}.$$

Dat wil zeggen dat de expected frequentie voor de cel linksboven, oftewel E_{11}, gelijk is aan

$$E_{11} = \frac{\text{rijtotaal}_1 \cdot \text{kolomtotaal}_1}{\text{totaal}} = \frac{200 \cdot 250}{500} = 100.$$

Het juiste antwoord is dus (d).

Opdracht 10.m4: Men wenst te toetsen of het aantal aanvragen bij een helpdesk van een computermaatschappij gelijk is verdeeld over de vijf werkdagen van de week. In een bepaalde week worden de volgende aantallen waarnemingen gedaan tussen maandag en vrijdag: 48, 65, 57, 72, 58. De *expected* frequentie voor het aantal aanvragen op een willekeurige maandag is dus ...

(a) 12

(b) 60

(c) 48

(d) 5

Uitwerking

In totaal zijn er in deze week $48+65+57+72+58 = 300$ waarnemingen gedaan. Indien het aantal aanvragen uniform verdeeld is over de vijf werkdagen, dan verwachten we dat 20% van de aanvragen op maandag wordt gedaan. De expected frequentie van het aantal aanvragen op maandag is dus gelijk aan $0,20 \cdot 300 = 60$ aanvragen. Het juiste antwoord is dus (b).

Opdracht 10.1: Bij een onderzoek naar de rookgewoonten van Nederlanders van 18 jaar en ouder werden door loting 200 proefpersonen gekozen die vervolgens werden ingedeeld naar leeftijd en naar rookgewoonte. De resultaten waren als volgt:

	Leeftijd			
	18− < 30	**30− < 45**	**45 en ouder**	**Totaal**
Roker	25	35	20	80
Niet-roker	55	25	40	120
Totaal	80	60	60	200

We gaan met behulp van de chikwadraattoets onderzoeken of de indelingen naar leeftijd en rookgewoonte al dan niet afhankelijk van elkaar zijn. We toetsen met $\alpha = 0,01$. De nulhypothese luidt: H_0 : onafhankelijkheid.

(a) Bereken de *expected*-tabel.

Uitwerking

Om de *expected*-tabel te bepalen, starten we vanuit een lege tabel waarin alleen de totalen (van de rijen en kolommen respectievelijk) gegeven zijn:

	Leeftijd			
	18− < 30	**30− < 45**	**45 en ouder**	**Totaal**
Roker				80
Niet-roker				120
Totaal	80	60	60	200

Voor iedere cel in de tabel bepalen we de expected frequentie met de formule:

$$E_{ij} = \frac{\text{rijtotaal}_i \cdot \text{kolomtotaal}_j}{\text{totaal}}$$

Dit geeft de volgende *expected*-tabel:

	Leeftijd			
	18− < 30	**30− < 45**	**45 en ouder**	**Totaal**
Roker	$\frac{80\cdot80}{200} = 32$	$\frac{80\cdot60}{200} = 24$	$\frac{80\cdot60}{200} = 24$	80
Niet-roker	$\frac{120\cdot80}{200} = 48$	$\frac{120\cdot60}{200} = 36$	$\frac{120\cdot60}{200} = 36$	120
Totaal	80	60	60	200

(b) Bereken de toetsingsgrootheid χ^2.

Uitwerking

De theoretische toetsingsgrootheid X^2 bepalen we aan de hand van de volgende formule:

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

waarbij $i = 1, 2$ de index van de rij is, en $j = 1, 2, 3$ de index van de kolom is. Dit geeft in dit specifieke geval een geobserveerde toetsingsgrootheid

$$\begin{aligned}
\chi^2 &= \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\
&= \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \ldots + \frac{(O_{23} - E_{23})^2}{E_{23}} \\
&= \frac{(25 - 32)^2}{32} + \frac{(35 - 24)^2}{24} + \ldots + \frac{(40 - 36)^2}{36} \\
&\approx 12,0660
\end{aligned}$$

(c) Hoeveel vrijheidsgraden heeft de chikwadraatverdeling die gebruikt moet worden?

Uitwerking

Bij een χ^2-toets voor onafhankelijkheid van twee nominale variabelen is het aantal vrijheidsgraden gelijk aan het aantal cellen waarvoor je een waarde vrij kunt kiezen. Dit is gelijk aan

$$\text{df} = (\#\text{rijen} - 1) \cdot (\#\text{kolommen} - 1) = (2 - 1) \cdot (3 - 1) = 2.$$

In deze hypothesetoets voor onafhankelijkheid heeft de toetsingsgrootheid een chikwadraatverdeling met df $= 2$ vrijheidsgraden.

(d) Geef het kritieke gebied aan van de grootheid X.

Uitwerking

Merk op dat onafhankelijkheid waarschijnlijker is als de toetsingsgrootheid dichter bij 0 ligt. Dit volgt uit de formule van de toetsingsgrootheid X, omdat in dat geval de observed frequenties dicht in de buurt van de expected frequenties liggen.
Het kritieke gebied is dus van de vorm $[g, \infty)$, waarbij g de grenswaarde is waarvoor geldt $P(X \geq g) = \alpha$. Dit kunnen we met de grafische rekenmachine bepalen aan de hand van

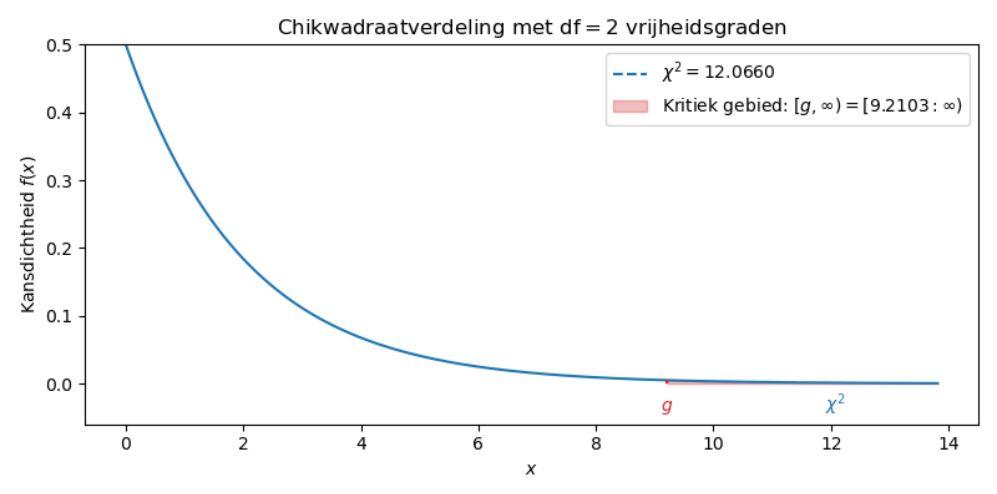
$$y_1 = \chi^2\text{cdf}(\text{lower} = x; \text{upper} = 10^{10}; \text{df} = 2)$$
$$y_2 = \alpha = 0,01$$

De numerieke solver geeft $x \approx 9,2103$. Dat betekent dat het kritieke gebied gelijk is aan $[9,2103; \infty)$.

(e) Wat is uw eindconclusie?

Uitwerking

De geobserveerde toetsingsgrootheid $\chi^2 \approx 12,0660$ ligt in het kritieke gebied, omdat $12,0660 > 9,2103$. Dit betekent dat de nulhypothese H_0 wordt verworpen. Er is voldoende bewijs om aan te nemen dat de twee nominale variabelen "leeftijd" en "rookgewoonte" afhankelijk zijn van elkaar.



Opdracht 10.5: Bij een onderzoek naar het gebruik van internet werden de respondenten onderverdeeld naar leeftijd en het wel of niet werken met internet. Voor 400 respondenten leverde dit de volgende tabel:

Internetgebruik

Leeftijd	**Wel internet**	**Geen internet**	**Totaal**
Tot en met 44 jaar	143	77	220
45 jaar en ouder	97	83	180
Totaal	240	160	400

(a) De vraag is of deze indelingen onafhankelijk zijn. Bereken de *expected*-tabel en voer de chikwadraattoets uit met behulp van de tabel (kies $\alpha = 0,01$).

Uitwerking

Bij een chikwadraattoets voor onafhankelijkheid beginnen we met het definiëren van de nulhypothese H_0 en de alternatieve hypothese H_1.

H_0 : "leeftijd" en "internetgebruik" zijn onafhankelijk van elkaar.

H_1 : "leeftijd" en "internetgebruik" zijn afhankelijk van elkaar.

Het significantieniveau $\alpha = 0,01$ is gegeven in de opdracht, net als de geobserveerde data. Om de toetsingsgrootheid X^2 te kunnen bepalen, moeten we eerst de *expected*-tabel berekenen. Hiervoor starten we vanuit een lege tabel waarin alleen de totalen (van de rijen en kolommen respectievelijk) gegeven zijn:

Internetgebruik			
Leeftijd	**Wel internet**	**Geen internet**	**Totaal**
Tot en met 44 jaar			220
45 jaar en ouder			180
Totaal	240	160	400

Voor iedere cel in de tabel bepalen we de expected frequentie met de formule:

$$E_{ij} = \frac{\text{rijtotaal}_i \cdot \text{kolomtotaal}_j}{\text{totaal}}$$

Dit geeft de volgende *expected*-tabel:

Internetgebruik			
Leeftijd	**Wel internet**	**Geen internet**	**Totaal**
Tot en met 44 jaar	$\frac{220 \cdot 240}{400} = 132$	$\frac{220 \cdot 160}{400} = 88$	220
45 jaar en ouder	$\frac{180 \cdot 240}{400} = 108$	$\frac{180 \cdot 160}{400} = 72$	180
Totaal	240	160	400

De toetsingsgrootheid X^2 bepalen we aan de hand van de volgende formule:

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

waarbij $i = 1, 2$ de index van de rij is, en $j = 1, 2$ de index van de kolom is. Dit geeft in dit specifieke geval een geobserveerde toetsingsgrootheid

$$\begin{aligned}
\chi^2 &= \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\
&= \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{22} - E_{22})^2}{E_{22}} \\
&= \frac{(143 - 132)^2}{132} + \frac{(77 - 88)^2}{88} + \frac{(97 - 108)^2}{108} + \frac{(83 - 72)^2}{72} \\
&\approx 4,6402
\end{aligned}$$

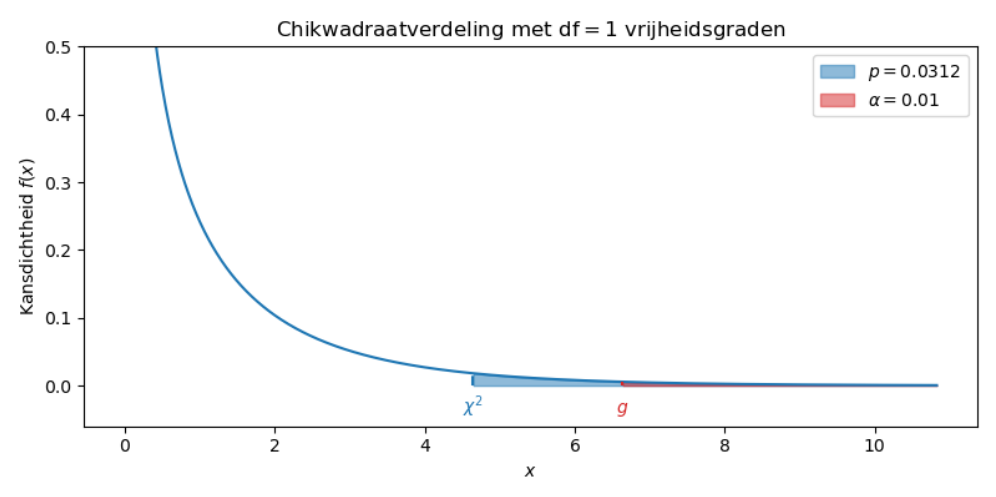
Onder de nulhypothese volgt de toetsingsgrootheid X een chikwadraatverdeling met

$$\text{df} = (\#\text{rijen} - 1) \cdot (\#\text{kolommen} - 1) = (2-1) \cdot (2-1) = 1$$

vrijheidsgraad. De p-waarde (rechteroverschrijdingskans) die hoort bij deze geobserveerde toetsingsgrootheid χ^2 is gelijk aan

$$\begin{aligned} p &= P(X^2 \geq \chi^2) \\ &= \chi^2\text{cdf}(\text{lower} = \chi^2 \approx 4,6402; \text{upper} = 10^{10}; \text{df} = 1) \\ &\approx 0,0312 \end{aligned}$$

Aangezien de p-waarde groter is dan het significantieniveau $\alpha = 0,01$, wordt H_0 niet verworpen. Er is onvoldoende bewijs om de hypothese te verwerpen dat de twee nominale variabelen "leeftijd" en "internetgebruik" onafhankelijk zijn van elkaar.



Opdracht 10.7: Bij een onderzoek naar de gevolgen van roken is gemeten hoe het gesteld is met de bloeddruk van 50-jarige mannen. In het onderzoek waren 200 mannen betrokken, waarvan 40 te kwalificeren zijn als stevige roker, 30 als gelegenheidsrokers en 130 als niet-rokers. Hun bloeddruk werd een van de volgende drie kwalificaties gegeven, namelijk normaal, licht verhoogd of ernstig verhoogd. Dat leverde de volgende tabel:

Bloeddruk	**Stevige roker**	**Matige roker**	**Niet-roker**
Normaal	6	8	86
Licht verhoogd	10	8	22
Ernstig verhoogd	24	14	22

Toets of rookgedrag en bloeddrukniveau significant samenhangen (kies $\alpha = 0,01$).

Uitwerking

Omdat we weer te maken hebben met twee *nominale* variabelen, namelijk "rookgedrag" en "bloeddrukniveau", kunnen we een chikwadraattoets uitvoeren om de onafhankelijkheid tussen beide variabelen te toetsen.
Bij een chikwadraattoets voor onafhankelijkheid beginnen we met het definiëren van

de nulhypothese H_0 en de alternatieve hypothese H_1.

H_0 : "rookgedrag" en "bloeddrukniveau" zijn onafhankelijk van elkaar.

H_1 : "rookgedrag" en "bloeddrukniveau" zijn afhankelijk van elkaar.

Het significantieniveau $\alpha = 0,01$ is gegeven in de opdracht, net als de geobserveerde data. Om de toetsingsgrootheid X^2 te kunnen bepalen, moeten we eerst de *expected*-tabel berekenen. Hiervoor starten we vanuit een lege tabel waarin alleen de totalen (van de rijen en kolommen respectievelijk) gegeven zijn:

Bloeddruk	**Stevige roker**	**Matige roker**	**Niet-roker**	**Totaal**
Normaal				100
Licht verhoogd				40
Ernstig verhoogd				60
Totaal	40	30	130	200

Voor iedere cel in de tabel bepalen we de expected frequentie met de formule:

$$E_{ij} = \frac{\text{rijtotaal}_i \cdot \text{kolomtotaal}_j}{\text{totaal}}$$

Dit geeft de volgende *expected*-tabel:

Bloeddruk	**Stevige roker**	**Matige roker**	**Niet-roker**	**Totaal**
Normaal	$\frac{40 \cdot 100}{200} = 20$	$\frac{30 \cdot 100}{200} = 15$	$\frac{130 \cdot 100}{200} = 65$	100
Licht verhoogd	$\frac{40 \cdot 40}{200} = 8$	$\frac{30 \cdot 40}{200} = 6$	$\frac{130 \cdot 40}{200} = 26$	40
Ernstig verhoogd	$\frac{40 \cdot 60}{200} = 12$	$\frac{30 \cdot 60}{200} = 9$	$\frac{130 \cdot 60}{200} = 39$	60
Totaal	40	30	130	200

De toetsingsgrootheid X^2 bepalen we aan de hand van de volgende formule:

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

waarbij $i = 1, 2, 3$ de index van de rij is, en $j = 1, 2, 3$ de index van de kolom is. Dit geeft in dit specifieke geval een geobserveerde toetsingsgrootheid

$$\begin{aligned}
\chi^2 &= \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\
&= \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \ldots + \frac{(O_{33} - E_{33})^2}{E_{33}} \\
&= \frac{(6 - 20)^2}{20} + \frac{(8 - 15)^2}{15} + \ldots + \frac{(22 - 39)^2}{39} \\
&\approx 43,8214
\end{aligned}$$

Onder de nulhypothese volgt de toetsingsgrootheid X een chikwadraatverdeling met

$$\text{df} = (\#\text{rijen} - 1) \cdot (\#\text{kolommen} - 1) = (3-1) \cdot (3-1) = 4$$

vrijheidsgraden. De p-waarde (rechteroverschrijdingskans) die hoort bij deze geobserveerde toetsingsgrootheid χ^2 is gelijk aan

$$\begin{aligned} p &= P(X^2 \geq \chi^2) \\ &= \chi^2\text{cdf}(\text{lower} = \chi^2 \approx 43,8214; \text{upper} = 10^{10}; \text{df} = 4) \\ &\approx 6,9879 \cdot 10^{-9} \end{aligned}$$

Aangezien de p-waarde extreem klein is, betekent dit dat de geobserveerde toetsingsgrootheid χ^2 extreem hoge waarde heeft onder de aanname van onafhankelijkheid. Omdat de p-waarde veel kleiner is dan het significantieniveau $\alpha = 0,01$, wordt H_0 verworpen. Er is voldoende bewijs om aan te nemen dat de twee nominale variabelen "rookgedrag" en "bloeddrukniveau" afhankelijk zijn van elkaar.

Side note: *de toetsuitslag van het verwerpen van H_0 laat niet zien hoe deze afhankelijkheid eruit ziet. Als je de ruwe data bekijkt, zie je echter dat onder de niet-rokers een groot deel gewoon een normale bloeddruk heeft, terwijl voor stevige rokers de meeste juist een ernstig verhoogde bloeddruk hebben. Een analyse van de data is dus nodig om te bekijken wat de daadwerkelijke samenhang is tussen de variabelen.*

Opdracht 10.11: In een ziekenhuis worden dagelijks vijf orthopedische operaties uitgevoerd. Bekend is dat deze in 20% van de gevallen leiden tot complicaties waardoor de patiënt enige tijd moet verblijven op de afdeling intensive care. Voor een periode van 100 dagen leidde dit tot de volgende aantallen verwijzingen naar de afdeling intensive care:

Aantal per dag (k)	**Aantal dagen met k verwijzingen**
0	15
1	25
2	36
3	12
4	10
5	2
Totaal	100 dagen

Toets met $\alpha = 0,05$ of de waargenomen verdeling overeenstemt met een binomiale verdeling met $p = 0,20$.

Uitwerking

Omdat we willen kijken of de gegeven data overeenkomen met een specifieke discrete kansverdeling, in dit geval de binomiale verdeling met $p = 0,20$, kunnen we een chikwadraattoets voor aanpassing (goodness-of-fit test) uitvoeren. Laat X een kansvariabele zijn die het aantal verwijzingen naar de afdeling intensive care telt op een willekeurige dag. Verder is gegeven dat er per dag 5 orthopedische operaties zijn, die in 20% van de gevallen leidt tot complicaties.
Bij een chikwadraattoets voor aanpassing beginnen we met het definiëren van de nulhypothese H_0 en de alternatieve hypothese H_1.

H_0 : X is binomiaal verdeeld met parameters $n = 5$ en $p = 0,20$.

H_1 : X is NIET binomiaal verdeeld met parameters $n = 5$ en $p = 0,20$.

Het significantieniveau $\alpha = 0,05$ is gegeven in de opdracht, net als de geobserveerde data. Om de toetsingsgrootheid X^2 te kunnen bepalen, moeten we eerst de *expected*-tabel berekenen. Omdat onder de nulhypothese H_0 geldt dat $X \sim$ Binomiaal$(n = 5; p = 0,20)$, kunnen we deze verwachte aantallen uitrekenen. Omdat we 100 afzonderlijke dagen bekijken, moet dit ook worden meegenomen in de verwachte frequenties.

Aantal per dag (k)	**Observed**	**Expected**
0	15	$100 \cdot \text{binompdf}(n = 5; p = 0,20; k = 0) = 32,768$
1	25	$100 \cdot \text{binompdf}(n = 5; p = 0,20; k = 1) = 40,96$
2	36	$100 \cdot \text{binompdf}(n = 5; p = 0,20; k = 2) = 20,48$
3	12	$100 \cdot \text{binompdf}(n = 5; p = 0,20; k = 3) = 5,12$
4	10	$100 \cdot \text{binompdf}(n = 5; p = 0,20; k = 4) = 0,64$
5	2	$100 - 32,768 - 40,96 - \ldots - 0,64 = 0,032$
Totaal	100 dagen	100 dagen

Merk op dat we de chikwadraatverdeling enkel konden gebruiken zodra de verwachte frequenties allemaal minstens 5 zijn. Om dit te bereiken, moeten we categorieën samennemen, namelijk 3 tot en met 5:

Aantal per dag (k)	**Observed**	**Expected**
0	15	$100 \cdot \text{binompdf}(n = 5; p = 0,20; k = 0) = 32,768$
1	25	$100 \cdot \text{binompdf}(n = 5; p = 0,20; k = 1) = 40,96$
2	36	$100 \cdot \text{binompdf}(n = 5; p = 0,20; k = 2) = 20,48$
≥ 3	24	$100 - 32,768 - 40,96 - 20,48 = 5,792$
Totaal	100 dagen	100 dagen

De toetsingsgrootheid X^2 bepalen we aan de hand van de volgende formule:

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i},$$

waarbij $i = 1, 2, 3, 4$ de index van de rij is. Dit geeft in dit specifieke geval een geobserveerde toetsingsgrootheid

$$\begin{aligned}
\chi^2 &= \sum_i \frac{(O_i - E_i)^2}{E_i} \\
&= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4} \\
&= \frac{(15 - 32,768)^2}{32,768} + \frac{(25 - 40,96)^2}{40,96} + \frac{(36 - 20,48)^2}{20,48} + \frac{(24 - 5,792)^2}{5,792} \\
&\approx 84,8540
\end{aligned}$$

Onder de nulhypothese volgt de toetsingsgrootheid X een chikwadraatverdeling met

$$\text{df} = (\#\text{rijen} - 1) = 4 - 1 = 3$$

vrijheidsgraden. De p-waarde (rechteroverschrijdingskans) die hoort bij deze geobserveerde toetsingsgrootheid χ^2 is gelijk aan

$$\begin{aligned} p &= P(X^2 \geq \chi^2) \\ &= \chi^2\text{cdf}(\text{lower} = \chi^2 \approx 84,8540; \text{upper} = 10^{10}; \text{df} = 3) \\ &\approx 2,7893 \cdot 10^{-18} \end{aligned}$$

Aangezien de p-waarde extreem klein is, betekent dit dat de geobserveerde toetsingsgrootheid χ^2 extreem hoge waarde heeft onder de aanname van een binomiale verdeling met $n = 5$ en $p = 0,20$. Omdat de p-waarde veel kleiner is dan het significantieniveau $\alpha = 0,01$, wordt H_0 verworpen. Er is voldoende bewijs om aan te nemen dat het aantal verwijzingen naar de afdeling intensive care niet binomiaal verdeeld is met parameters $n = 5$ en $p = 0,20$.

Side note: *eigenlijk moet je nog doordenken over wat de toetsuitslag betekent. Merk op dat volgens de metingen het totaal aantal doorverwijzingen in* 100 *dagen gelijk is aan*

$$0 \cdot 15 + 1 \cdot 25 + 2 \cdot 36 + 3 \cdot 12 + 4 \cdot 10 + 5 \cdot 2 = 183$$

Aangezien er elke dag 5 *operaties plaatsvinden, is het totaal aantal operaties gelijk aan* 500*. Dit betekent dat de fractie doorverwezen patiënten gelijk is aan* $\frac{183}{500} = 0,366$*. Deze fractie is veel groter dan* $p = 0,20$*, dus misschien is het niet zo realistisch om te verwachten dat de nulhypothese gaat kloppen. Een analyse van de ruwe data is dus nodig om te bekijken wat de daadwerkelijke samenhang is tussen de variabelen.*

Opdracht 10.12: Soms wil men voor een getrokken steekproef beoordelen of die als representatief mag worden beschouwd met betrekking tot een bepaald kenmerk of een bepaalde variabele. Met de chikwadraattoets voor aanpassing kan worden getoetst of de waargenomen verdeling in dit opzicht voldoende gelijkenis vertoont met de populatieopbouw. Bij een opinieonderzoek over de toekomst van Europa worden 480 kiesgerechtigde Nederlanders ondervraagd. Van alle ondervraagden is het opleidingsniveau genoteerd. In de volgende tabel wordt dit opleidingsniveau vergeleken met de totale Nederlandse bevolking:

	Laag	**Matig**	**Redelijk**	**Hoog**	**Totaal**
Steekproef	134	144	129	73	480
Populatie	28%	36%	24%	12%	100%

Toets met $\alpha = 0,05$ of de steekproef als representatief mag worden beschouwd.

Uitwerking

In deze opgave willen we kijken of de steekproef representatief is of niet. Hierbij horen de volgende nulhypothese H_0 en de alternatieve hypothese H_1:

H_0 : de gekozen steekproef is representatief voor de populatie
H_1 : de gekozen steekproef is NIET representatief voor de populatie

Verder is het significantieniveau $\alpha = 0,05$ en de steekproefdata (zie tabel) gegeven. Om te toetsen of de gegeven steekproefdata (observed) overeenkomen representatief zijn, moeten we allereerst de verwachte frequenties berekenen gebaseerd op de populatiepercentages. Dit kunnen we doen door de populatiepercentages om te rekenen naar verwachte frequenties op basis van het kiezen van 480 kiesgerechtigde Nederlanders:

Opleidingsniveau	**Steekproeffrequenties (observed)**	**Verwachte frequenties (expected)**
Laag	134	$28\% \cdot 480 = 134,4$
Matig	144	$36\% \cdot 480 = 172,8$
Redelijk	129	$24\% \cdot 480 = 115,2$
Hoog	73	$12\% \cdot 480 = 57,6$
Totaal	480	480

Omdat we willen kijken of de geobserveerde data overeenkomen met de verwachte frequenties (van een nominale variabele), gaan we een chikwadraattoets voor aanpassing (goodness-of-fit test) uitvoeren. Laat X een kansvariabele zijn die het aantal verwijzingen naar de afdeling intensive care telt op een willekeurige dag. Verder is gegeven dat er per dag 5 orthopedische operaties zijn, die in 20% van de gevallen leidt tot complicaties. Bij een chikwadraattoets voor aanpassing beginnen we met het definiëren van de nulhypothese H_0 en de alternatieve hypothese H_1. De verwachte frequenties zijn allemaal groter dan of gelijk aan 5, dus kunnen we de toetsingsgrootheid direct berekenen. Merk op dat we de chikwadraatverdeling enkel konden gebruiken zodra de verwachte frequenties allemaal minstens 5 zijn.
De toetsingsgrootheid X^2 bepalen we aan de hand van de volgende formule:

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i},$$

waarbij $i = 1, 2, 3, 4$ de index van de kolom is. Dit geeft in dit specifieke geval een geobserveerde toetsingsgrootheid

$$\begin{aligned}
\chi^2 &= \sum_i \frac{(O_i - E_i)^2}{E_i} \\
&= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4} \\
&= \frac{(134 - 134,4)^2}{134,4} + \frac{(144 - 172,8)^2}{172,8} + \frac{(129 - 115,2)^2}{115,2} + \frac{(73 - 57,6)^2}{57,6} \\
&\approx 10,5717
\end{aligned}$$

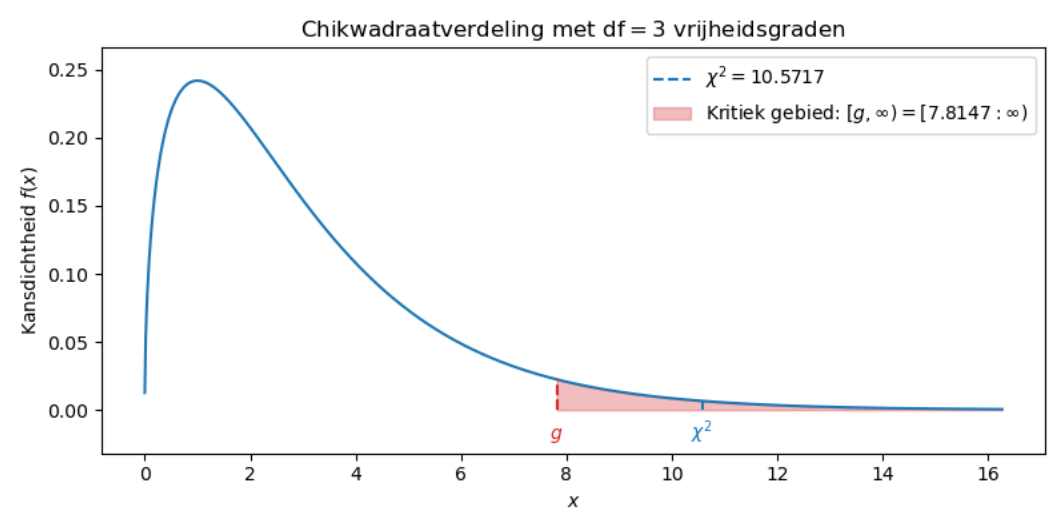
Onder de nulhypothese volgt de toetsingsgrootheid X^2 een chikwadraatverdeling met

$$\text{df} = (\#\text{categorieën} - 1) = 4 - 1 = 3$$

vrijheidsgraden. Extreem grote waarde van de toetsingsgrootheid duiden op grote verschillen tussen observed en expected frequenties, dus het kritieke gebied is van de vorm $[g, \infty)$. Deze grenswaarde g is de oplossing van de vergelijking

$$\chi^2\text{cdf}(\text{lower} = g; \text{upper} = 10^{10}; \text{df} = 3) = \alpha = 0,05$$

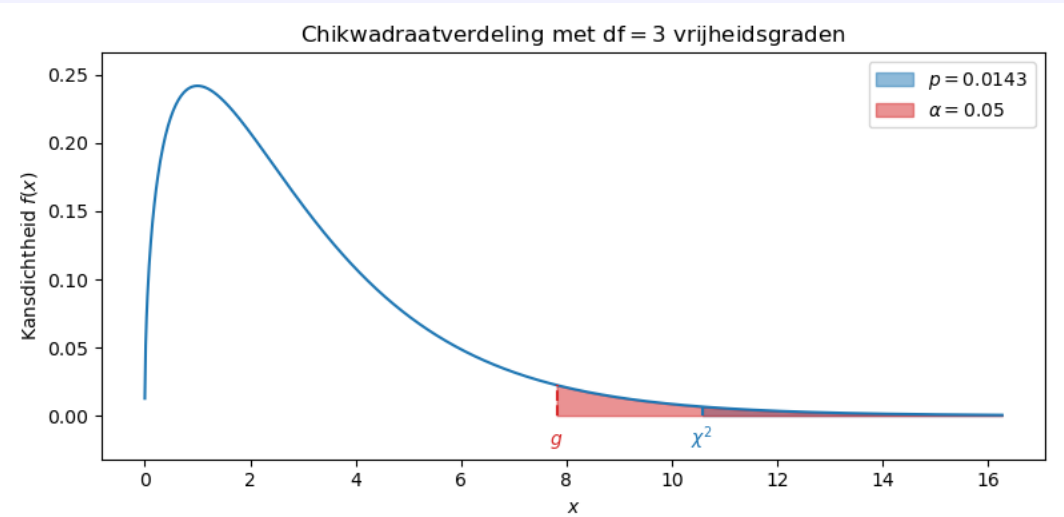
De solver optie geeft $g \approx 7,8147$. Aangezien de toetsingsgrootheid χ^2 groter dan deze grenswaarde is ($10,5717 > 7,8147$), geldt dat de toetsingsgrootheid in het kritieke gebied ligt en dus H_0 wordt verworpen. Er is voldoende bewijs om aan te nemen dat de steekproef niet representatief is voor de populatie kiesgerechtigde Nederlanders.



Alternatief: de p-waarde (rechteroverschrijdingskans) die hoort bij deze geobserveerde toetsingsgrootheid χ^2 is gelijk aan

$$\begin{aligned} p &= P(X^2 \geq \chi^2) \\ &= \chi^2\text{cdf}(\text{lower} = \chi^2 \approx 10,5717; \text{upper} = 10^{10}; \text{df} = 3) \\ &\approx 0,0143 \end{aligned}$$

Aangezien de p-waarde kleiner is dan het significantieniveau α, betekent dit dat de geobserveerde toetsingsgrootheid χ^2 een extreem hoge waarde heeft onder de aanname dat de steekproef representatief zou zijn voor de populatie. Dit betekent dat de nulhypothese H_0 wordt verworpen. Er is voldoende bewijs om aan te nemen dat de steekproef niet representatief is voor de populatie kiesgerechtigde Nederlanders.



Side note: *eigenlijk moet je nog doordenken over wat de toetsuitslag betekent. Merk op dat volgens de metingen het aantal respondenten met redelijk en hoog opleidingsniveau (veel) groter is dan verwacht, terwijl dit voor mensen met een matig opleidingsniveau een stuk lager dan verwacht is.*

Opdracht 10.13: Het aantal brandmeldingen dat in een stad per week is geregistreerd gedurende een periode van 100 weken blijkt uit de tabel. Toets of het aantal branden per week is te beschouwen als een kansvariabele die een Poissonverdeling volgt met $\mu = 1$ (kies $\alpha = 0,05$).

Aantal branden per week	**Frequentie**
0	48
1	24
2	16
3	8
4 of meer	4
Totaal	100 weken

Uitwerking

Omdat we willen kijken of de gegeven data overeenkomen met een specifieke discrete kansverdeling, in dit geval de Poissonverdeling met $\mu = 1$, kunnen we een chikwadraattoets voor aanpassing (goodness-of-fit test) uitvoeren. Laat X een kansvariabele zijn die het aantal branden telt in een willekeurige week. Verder is gegeven dat over een periode van 100 weken het aantal branden is geteld. Bij een chikwadraattoets voor aanpassing beginnen we met het definiëren van de nulhypothese H_0 en de alternatieve hypothese H_1.

H_0 : X is Poisson verdeeld met gemiddelde $\mu = 1$.
H_1 : X is NIET Poisson verdeeld met gemiddelde $\mu = 1$.

Het significantieniveau $\alpha = 0,05$ is gegeven in de opdracht, net als de geobserveerde data. Om de toetsingsgrootheid X^2 te kunnen bepalen, moeten we eerst de *expected*-tabel berekenen. Omdat onder de nulhypothese H_0 geldt dat $X \sim \text{Poisson}(\mu = 1)$, kunnen we deze verwachte aantallen uitrekenen. Omdat we 100 afzonderlijke weken bekijken, moet dit ook worden meegenomen in de verwachte frequenties.

Aantal branden per week	**Frequentie (observed)**	**Frequentie (expected)**
0	48	$100 \cdot \text{poissonpdf}(\mu = 1; k = 0) = 36,7879$
1	24	$100 \cdot \text{poissonpdf}(\mu = 1; k = 1) = 36,7879$
2	16	$100 \cdot \text{poissonpdf}(\mu = 1; k = 2) = 18,3940$
3	8	$100 \cdot \text{poissonpdf}(\mu = 1; k = 3) = 6,1313$
4 of meer	4	$100 - 36,7879 - 36,7879 - \ldots = 1,8988$
Totaal	100 dagen	100 dagen

Merk op dat we de chikwadraatverdeling enkel konden gebruiken zodra de verwachte frequenties allemaal minstens 5 zijn. Om dit te bereiken, moeten we categorieën samenvoegen, namelijk 3 en 4 of meer:

Aantal branden per week	**Frequentie (observed)**	**Frequentie (expected)**
0	48	$36,7879$
1	24	$36,7879$
2	16	$18,3940$
3 of meer	12	$8,0301$
Totaal	100 dagen	100 dagen

De toetsingsgrootheid X^2 bepalen we aan de hand van de volgende formule:

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i},$$

waarbij $i = 1, 2, 3, 4$ de index van de rij is. Dit geeft in dit specifieke geval een geobserveerde toetsingsgrootheid

$$\begin{aligned}
\chi^2 &= \sum_i \frac{(O_i - E_i)^2}{E_i} \\
&= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4} \\
&= \frac{(48 - 36,7879)^2}{36,7879} + \frac{(24 - 36,7879)^2}{36,7879} + \frac{(16 - 18,3940)^2}{18,3940} + \frac{(12 - 8,0301)^2}{8,0301} \\
&\approx 10,1366
\end{aligned}$$

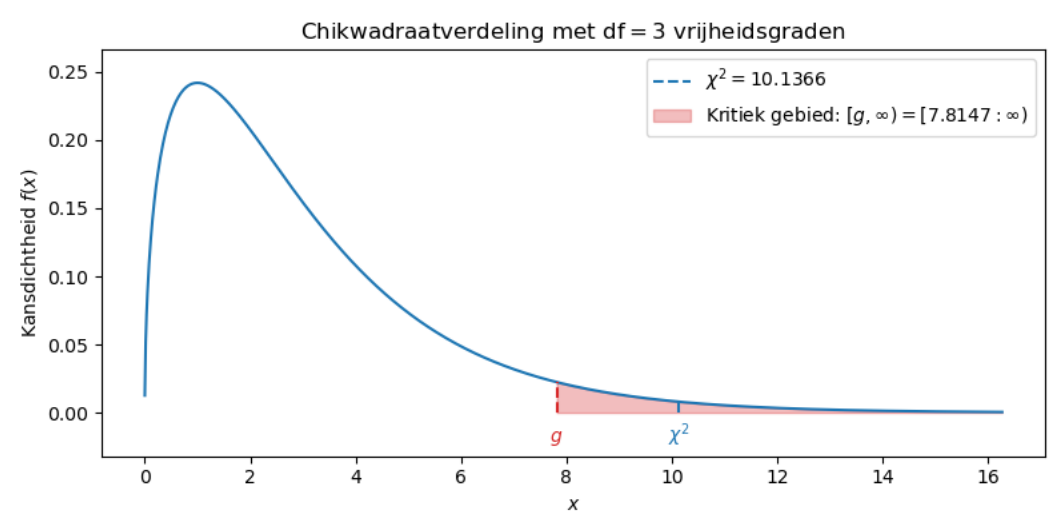
Onder de nulhypothese volgt de toetsingsgrootheid X een chikwadraatverdeling met

$$\text{df} = (\#\text{categorieën} - 1) = 4 - 1 = 3$$

vrijheidsgraden.
Extreem grote waarde van de toetsingsgrootheid duiden op grote verschillen tussen observed en expected frequenties, dus het kritieke gebied is van de vorm $[g, \infty)$. Deze grenswaarde g is de oplossing van de vergelijking

$$\chi^2\text{cdf}(\text{lower} = g; \text{upper} = 10^{10}; \text{df} = 3) = \alpha = 0,05$$

De solver optie geeft $g \approx 7,8147$. Aangezien de toetsingsgrootheid χ^2 groter dan deze grenswaarde is ($10,1366 > 7,8147$), geldt dat de toetsingsgrootheid in het kritieke gebied ligt en dus H_0 wordt verworpen. Er is voldoende bewijs om aan te nemen dat het aantal branden per week niet een Poissonverdeling met gemiddelde $\mu = 1$ volgt.



Alternatief: de p-waarde (rechteroverschrijdingskans) die hoort bij deze geobserveerde toetsingsgrootheid χ^2 is gelijk aan

$$\begin{aligned} p &= P(X^2 \geq \chi^2) \\ &= \chi^2\text{cdf}(\text{lower} = \chi^2 \approx 10,1366; \text{upper} = 10^{10}; \text{df} = 3) \\ &\approx 0.0174 \end{aligned}$$

Aangezien de p-waarde kleiner is dan het significantieniveau α, betekent dit dat de geobserveerde toetsingsgrootheid χ^2 een extreem hoge waarde heeft onder de aanname van een Poissonverdeling met $\mu = 1$. Dit betekent dat op basis van deze steekproef de nulhypothese H_0 wordt verworpen. Er is voldoende bewijs om aan te nemen dat het aantal branden per week niet een Poissonverdeling met gemiddelde $\mu = 1$ volgt.