Defensie Ondersteuningscommando
Ministerie van Defensie

Statistiek: college 10

De chikwadraatverdeling

DOSCO

Dr. ir. Danny Blom
Nederlandse Defensie Academie
Faculteit Militaire Wetenschappen

juli 2025

Terugblik

- Punt- en intervalschatters
- Betrouwbaarheid- en voorspellingsintervallen
- De methode van hypothesetoetsen
- Hypothesetoetsen voor het gemiddelde van de normale verdeling

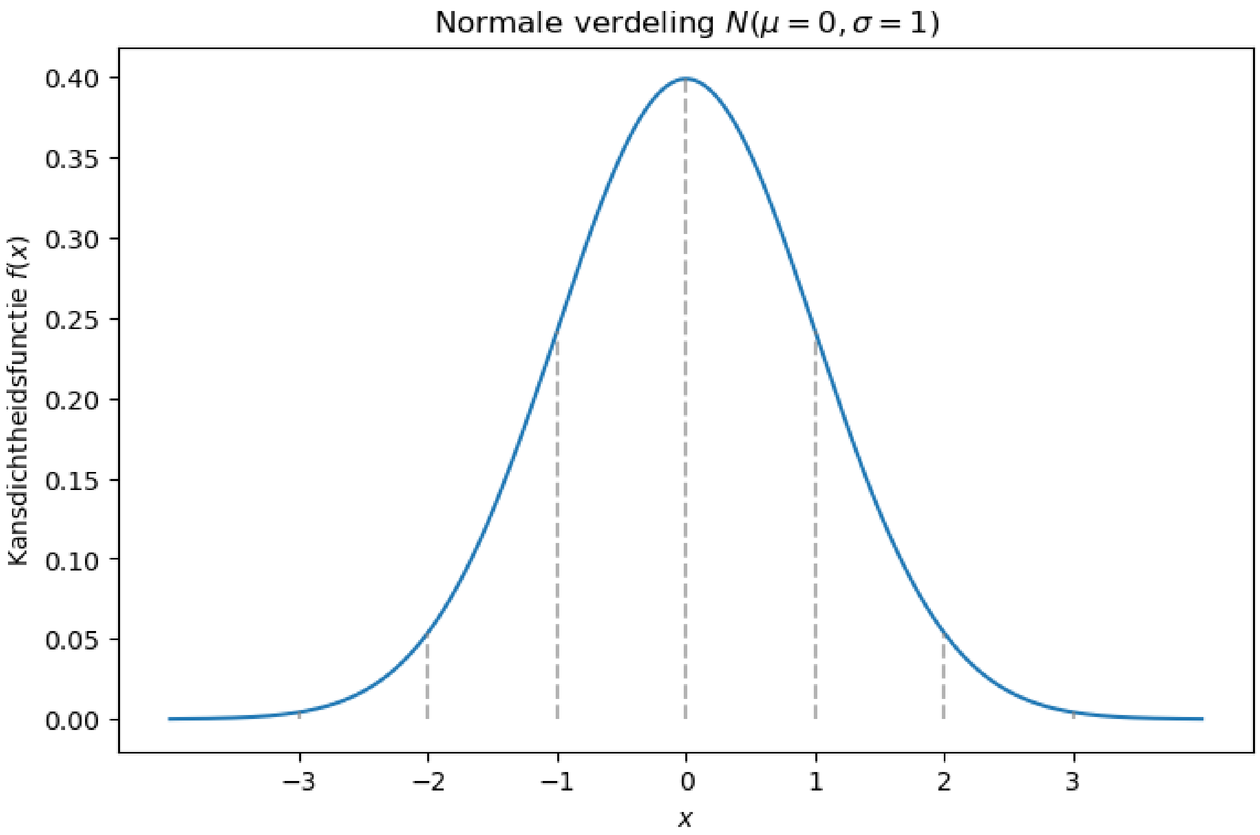
Leerdoelen

Aan het eind van dit college kunnen studenten:

- De eigenschappen van de chikwadraatverdeling benoemen en uitleggen
- De verschillende situaties kunnen benoemen waarin de chikwadraatverdeling een rol speelt.
- Met behulp van de chikwadraatverdeling een toets van onafhankelijkheid tussen twee nominale variabelen uitvoeren.
- De chikwadraatverdeling gebruiken om een aanpassingstoets voor discrete kansverdelingen (goodness-of-fit test) uit te voeren.

Recap: de standaardnormale verdeling $Z \sim N(0,1)$

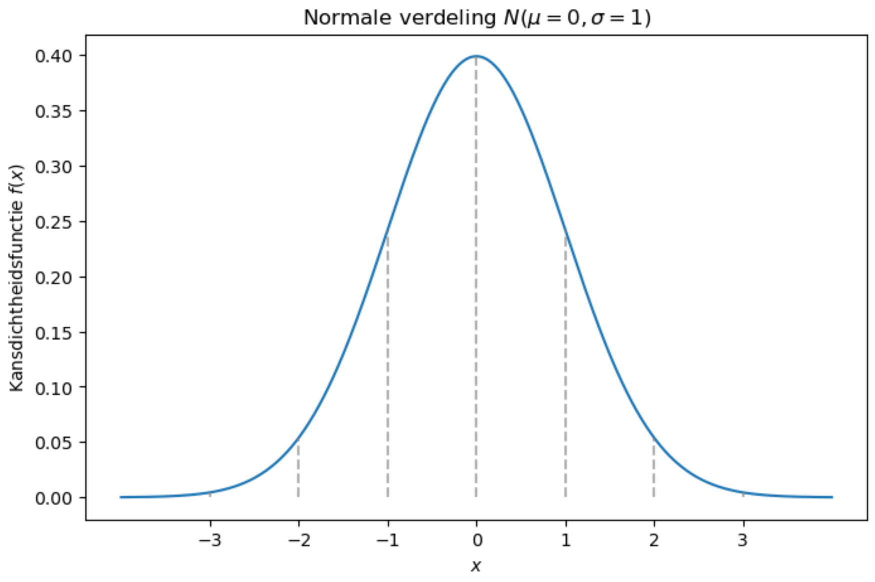
- Gemiddelde $\mu = 0$ en standaardafwijking $\sigma = 1$
- Symmetrisch rond $x = 0$
- Belkromme
- Standaardisatie van normale verdeling



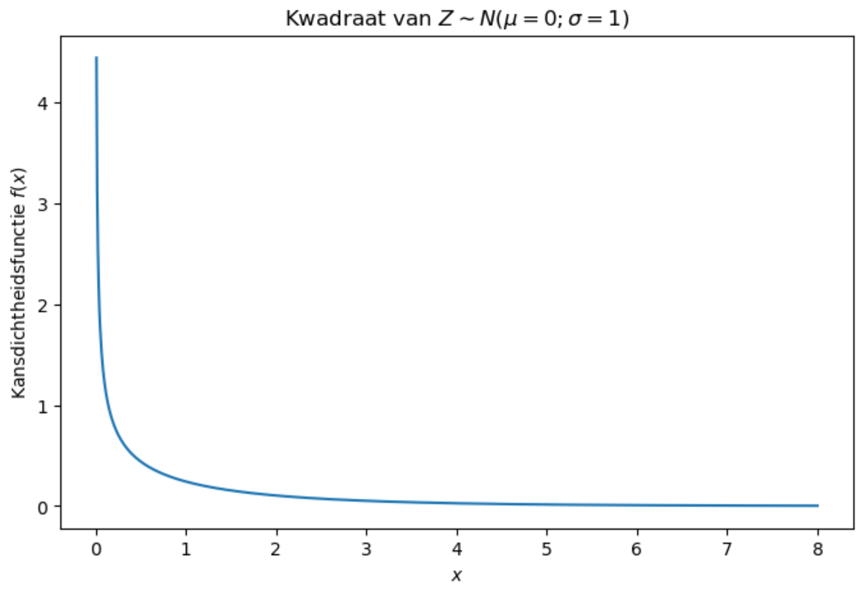
Wat kunnen we zeggen over het kwadraat Z^2 van $Z \sim N(0,1)$?

- $Z^2 \geq 0$: waarden zijn niet-negatief
- Niet meer symmetrisch
- Bouwsteen voor de chikwadraatverdeling

Kansdichtheidsfunctie van Z



Kansdichtheidsfunctie van Z^2



De chikwadraatverdeling

De **chikwadraatverdeling** is een continue kansverdeling die hoort bij de som van n kwadraten van standaardnormaal verdeelde kansvariabelen:

$$X = Z_1^2 + Z_2^2 + \cdots + Z_n^2 \sim \chi_n^2$$

Eigenschappen:

- We noemen het getal n ook wel het aantal **vrijheidsgraden** van de chikwadraatverdeling
- $E[X] = n$
- $\mathrm{Var}(X) = 2n$

Merk op: de formule voor de (steekproef)variantie bevat ook een som van kwadraten!

De chikwadraatverdeling

interactive-chisq.streamlit.app

Toepassing 1: chikwadraattoets voor onafhankelijkheid

Stel dat we willen onderzoeken onder de gewonde soldaten tijdens een oorlogsmissie of het type verwonding samenhangt met hun taak tijdens de missie.

- Interactie tussen twee **categorische** variabelen, oftewel **nominaal** of **ordinaal**

Vraag: zijn deze twee variabelen (type verwonding / taakstelling) onafhankelijk van elkaar?

Stap 1: formuleer de nulhypothese H_0 en alternatieve hypothese H_1

Nulhypothese:

H_0: de variabelen “type verwonding” en “taakstelling” zijn onafhankelijke variabelen.

Alternatieve hypothese:

H_1: de variabelen “type verwonding” en “taakstelling” zijn afhankelijke variabelen.

Stap 2: bepaal de kans α op een type-I fout

We kunnen zelf een kans α op een type-I fout kiezen

- Hoeveel risico wil je nemen om H_0 (onafhankelijkheid van de twee variabelen) onterecht te verwerpen?

Kans α op een type-I fout: $\alpha = 0{,}05$

- Indien de twee variabelen onafhankelijk zijn (H_0 is waar), zal deze in 95% van alle steekproefresultaten ook daadwerkelijk worden aangenomen.

Stap 3: verzamelen van data

Stel dat gedurende een jaar tijdens verschillende missies deze aantal worden geregistreerd:

		Taakstelling			
		Patrouille	Mijnruiming	Logistiek	**Totaal**
Type verwonding	Hoofd	12	18	30	**60**
	Romp	8	22	20	**50**
	Ledematen	10	10	20	**40**
	Totaal	**30**	**50**	**70**	**150**

Stap 4: bepalen van de toetsingsgrootheid X^2

Om te toetsen of het type verwonding onafhankelijk is van de taak die wordt uitgevoerd, moeten we de verzamelde data vergelijken met de verwachte uitkomsten.

Wat voor uitkomsten verwachten we bij onafhankelijkheid?

		Taakstelling			
		Patrouille	Mijnruiming	Logistiek	**Totaal**
Type verwonding	Hoofd				**60**
	Romp				**50**
	Ledematen				**40**
	Totaal	**30**	**50**	**70**	**150**

Stap 4: bepalen van de toetsingsgrootheid X^2

Wat voor uitkomsten verwachten we bij onafhankelijkheid?

- Per type verwonding is het aantal gewonden ook evenredig verdeeld (verhouding 30 : 50 : 70)
- Algemeen: de **verwachte frequentie** is $E_{ij} = \frac{\text{rijtotaal}_i \cdot \text{kolomtotaal}_j}{\text{totaal}}$

		Taakstelling			
		Patrouille	Mijnruiming	Logistiek	**Totaal**
Type verwonding	Hoofd	$\frac{30 \cdot 60}{150} = 12$	$\frac{50 \cdot 60}{150} = 20$	$\frac{70 \cdot 60}{150} = 28$	**60**
	Romp	$\frac{30 \cdot 50}{150} = 10$	$\frac{50 \cdot 50}{150} = 16{,}66$	$\frac{70 \cdot 50}{150} = 23{,}33$	**50**
	Ledematen	$\frac{30 \cdot 40}{150} = 8$	$\frac{50 \cdot 40}{150} = 13{,}33$	$\frac{70 \cdot 40}{150} = 18{,}67$	**40**
	Totaal	**30**	**50**	**70**	**150**

Stap 4: bepalen van de toetsingsgrootheid X^2

Geobserveerde frequenties uit data ("observed")

		Taakstelling			
		Patrouille	Mijnruiming	Logistiek	**Totaal**
Type verwonding	Hoofd	12	18	30	**60**
	Romp	8	22	20	**50**
	Ledematen	10	10	20	**40**
	Totaal	**30**	**50**	**70**	**150**

Verwachte frequenties ("expected")

		Taakstelling			
		Patrouille	Mijnruiming	Logistiek	**Totaal**
Type verwonding	Hoofd	12	20	28	**60**
	Romp	10	16,6666	23,3333	**50**
	Ledematen	8	13,3333	18,6666	**40**
	Totaal	**30**	**50**	**70**	**150**

Toetsingsgrootheid: $X^2 = \sum_{i,j} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2(\text{df})$

- O_{ij} (E_{ij}): geobserveerde (verwachte) aantallen voor rij i, kolom j
- $\text{df} = (\#\text{rijen} - 1) \cdot (\#\text{kolommen} - 1) = (3-1) \cdot (3-1) = 4$ vrijheidsgraden

Stap 4: bepalen van de toetsingsgrootheid X^2

Geobserveerde frequenties uit data ("observed")

		Taakstelling			
		Patrouille	Mijnruiming	Logistiek	**Totaal**
Type verwonding	Hoofd	12	18	30	**60**
	Romp	8	22	20	**50**
	Ledematen	10	10	20	**40**
	Totaal	**30**	**50**	**70**	**150**

Verwachte frequenties ("expected")

		Taakstelling			
		Patrouille	Mijnruiming	Logistiek	**Totaal**
Type verwonding	Hoofd	12	20	28	**60**
	Romp	10	16,6666	23,3333	**50**
	Ledematen	8	13,3333	18,6666	**40**
	Totaal	**30**	**50**	**70**	**150**

De **geobserveerde toetsingsgrootheid** χ^2 is dan gelijk aan

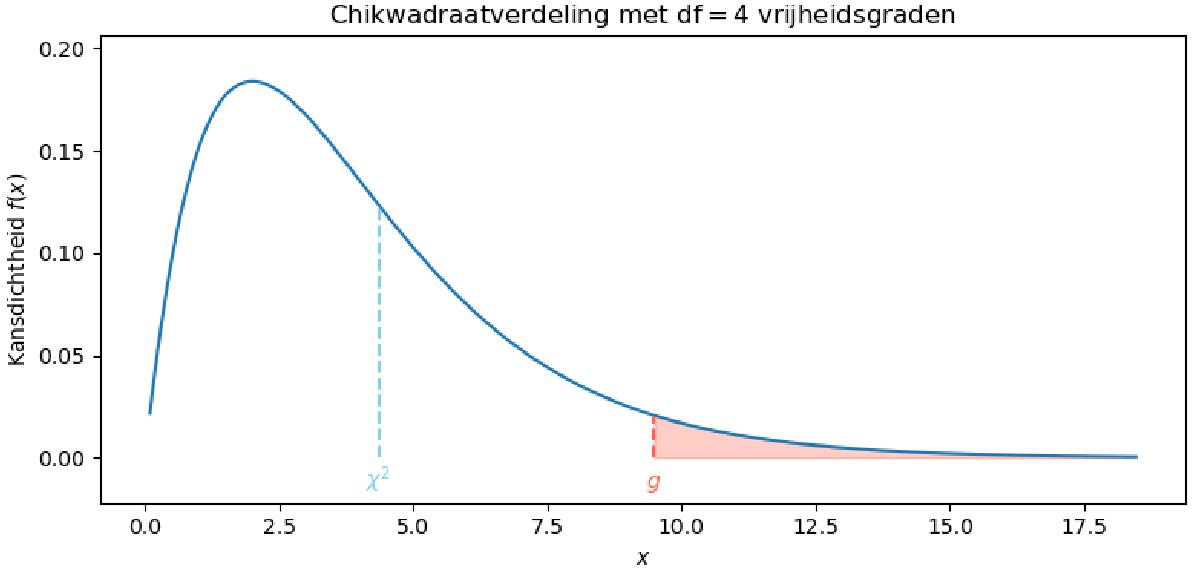
$$\chi^2 = \frac{(12-12)^2}{12} + \frac{(18-20)^2}{20} + \cdots + \frac{(20-18{,}6666)^2}{18{,}6666} \approx 4{,}3543$$

Stap 5: geef een conclusie en formuleer in de originele context

Methode 1 (kritiek gebied): het **kritieke gebied** wordt gegeven door $[g, \infty)$, waarbij de grens g de oplossing is van $\chi^2\text{cdf}(a = X; b = 10^{99}; df = 4) = \alpha = 0{,}05$

Met de solver optie krijgen we $g = 9{,}4877$

$\rightarrow \chi^2 \approx 4{,}3543$ ligt niet in het kritieke gebied
$\rightarrow H_0$ wordt niet verworpen!

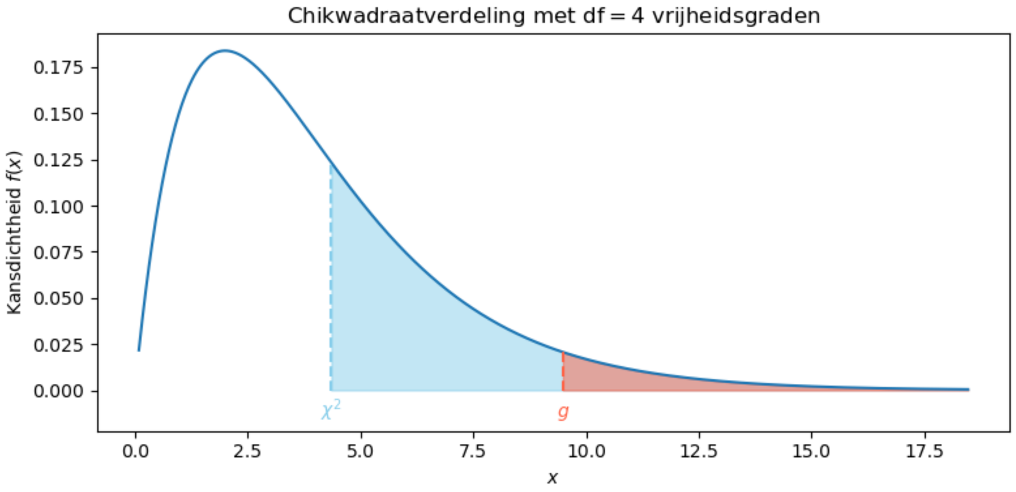


Er is onvoldoende reden om de onafhankelijkheid van de twee nominale variabelen "type verwonding" en "taakstelling" te verwerpen.

Stap 5: geef een conclusie en formuleer in de originele context

Methode 2 (p-waarde): de p-waarde (rechteroverschrijdingskans) wordt gegeven door

$$p = P(X^2 \geq \chi^2 \approx 4{,}3543\ ) = \chi^2\text{cdf}(a = 4{,}3543; b = 10^{99}; df = 4) \approx 0{,}3602$$



→ de p-waarde is groter dan het significantieniveau $\alpha = 0{,}05$, dus H_0 wordt niet verworpen.

Er is onvoldoende reden om de onafhankelijkheid van de twee nominale variabelen "type verwonding" en "taakstelling" te verwerpen.

Toepassing 2: aanpassingstoets (“goodness-of-fit test”)

De commandant van Paresto wil inzicht krijgen in de drukte tijdens piekuren in de bedrijfskantines. Hij wil graag weten of het aantal personen dat in een minuut binnenkomt tijdens de lunchpauze een Poissonverdeling met gemiddelde $\mu = 3{,}7$ personen.

Doel: bestuderen in hoeverre data overeenkomen met een **gegeven kansverdeling**

Stap 1: formuleer de nulhypothese H_0 en de alternatieve hypothese H_1

Nulhypothese $\boldsymbol{H_0}$:

Het aantal personen dat in een minuut de kantine binnenkomt volgt een Poissonverdeling met gemiddelde $\mu = 3{,}7$.

Alternatieve hypothese $\boldsymbol{H_1}$:

Het aantal personen dat in een minuut de kantine binnenkomt volgt **NIET** een Poissonverdeling met gemiddelde $\mu = 3{,}7$.

Stap 2: bepaal de kans α op een type-I fout

We kunnen zelf een kans α op een type-I fout kiezen

- Hoeveel risico wil je nemen om H_0 (de aanname van de $\text{Poisson}(\mu = 3{,}7)$ verdeling) onterecht te verwerpen?

Kans α op een type-I fout: $\alpha = 0{,}05$

- 5% kans op type-I fout, oftewel onterecht verwerpen dat het aantal binnenkomende mensen $\text{Poisson}(\mu = 3{,}7)$ verdeeld is

- Indien het aantal binnenkomende mensen inderdaad $\text{Poisson}(\mu = 3{,}7)$ verdeeld is (H_0 is waar), zal de nulhypothese in 95% van alle steekproefresultaten ook daadwerkelijk worden aangenomen.

Stap 3: verzamelen van data

We verzamelen een jaar lang elke dag een datapunt, waarmee we de volgende frequentietabel kunnen maken:

Aantal personen	**Geobserveerde frequenties O_i**
0	10
1	25
2	60
3	85
4	80
5	55
6	30
≥ 7	20
Totaal	**365**

Stap 3: verzamelen van data

We gaan opnieuw de geobserveerde frequenties vergelijken met de verwachte frequenties onder H_0. Merk op H_0 uitgaat van een Poissonverdeling met gemiddelde $\mu = 3{,}7$.

Aantal personen	Geobserveerde frequenties O_i	Verwachte frequenties E_i
0	10	$365 * \text{poissonpdf}(\lambda = 3{,}7; k = 0) \approx 9{,}0241$
1	25	$365 * \text{poissonpdf}(\lambda = 3{,}7; k = 1) \approx 33.3891$
2	60	$365 * \text{poissonpdf}(\lambda = 3{,}7; k = 2) \approx 61{,}7699$
3	85	$365 * \text{poissonpdf}(\lambda = 3{,}7; k = 3) \approx 76{,}1828$
4	80	$365 * \text{poissonpdf}(\lambda = 3{,}7; k = 4) \approx 70{,}4691$
5	55	$365 * \text{poissonpdf}(\lambda = 3{,}7; k = 5) \approx 52{,}1472$
6	30	$365 * \text{poissonpdf}(\lambda = 3{,}7; k = 6) \approx 32{,}1574$
≥ 7	20	$365 - 9{,}0241 - 33{,}3891 - \cdots \approx 29{,}8603$
Totaal	**365**	**365**

Stap 4: bepaal de toetsingsgrootheid X^2

Toetsingsgrootheid: $X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(\mathrm{df})$

- $\mathrm{df} = \#\text{categorieën} - 1 = 7$

Aantal personen	Geobserveerde frequenties O_i	Verwachte frequenties E_i
0	10	$365 * \mathrm{poissonpdf}(\lambda = 3{,}7; k = 0) \approx 9{,}0241$
1	25	$365 * \mathrm{poissonpdf}(\lambda = 3{,}7; k = 1) \approx 33.3891$
2	60	$365 * \mathrm{poissonpdf}(\lambda = 3{,}7; k = 2) \approx 61{,}7699$
3	85	$365 * \mathrm{poissonpdf}(\lambda = 3{,}7; k = 3) \approx 76{,}1828$
4	80	$365 * \mathrm{poissonpdf}(\lambda = 3{,}7; k = 4) \approx 70{,}4691$
5	55	$365 * \mathrm{poissonpdf}(\lambda = 3{,}7; k = 5) \approx 52{,}1472$
6	30	$365 * \mathrm{poissonpdf}(\lambda = 3{,}7; k = 6) \approx 32{,}1574$
≥ 7	20	$365 - 9{,}0241 - 33{,}3891 - \cdots \approx 29{,}8603$
Totaal	**365**	**365**

De geobserveerde toetsingsgrootheid χ^2 is dan

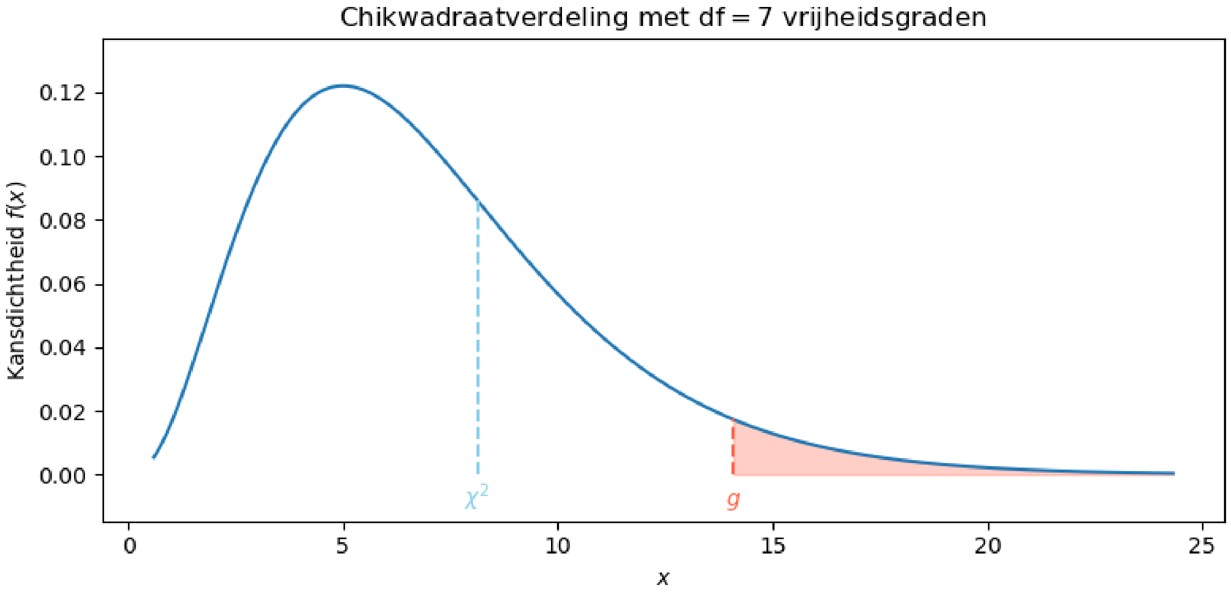
$$\chi^2 = \frac{(10 - 9{,}0241)^2}{9{,}0241} + \frac{(25 - 33{,}3891)^2}{33{,}3891} + \cdots + \frac{(20 - 29{,}8603)^2}{29{,}8603} \approx 8{,}1304$$

Stap 5: geef een conclusie en formuleer deze in de originele context

Methode 1 (kritiek gebied): het **kritieke gebied** wordt gegeven door $[g, \infty)$, waarbij de grens g de oplossing is van $\chi^2\mathrm{cdf}(a = X; b = 10^{99}; df = 7) = 0{,}05$

Met de solver optie krijgen we $g = 14{,}0671$

→ χ^2 ligt niet in het kritieke gebied
→ H_0 wordt niet verworpen!

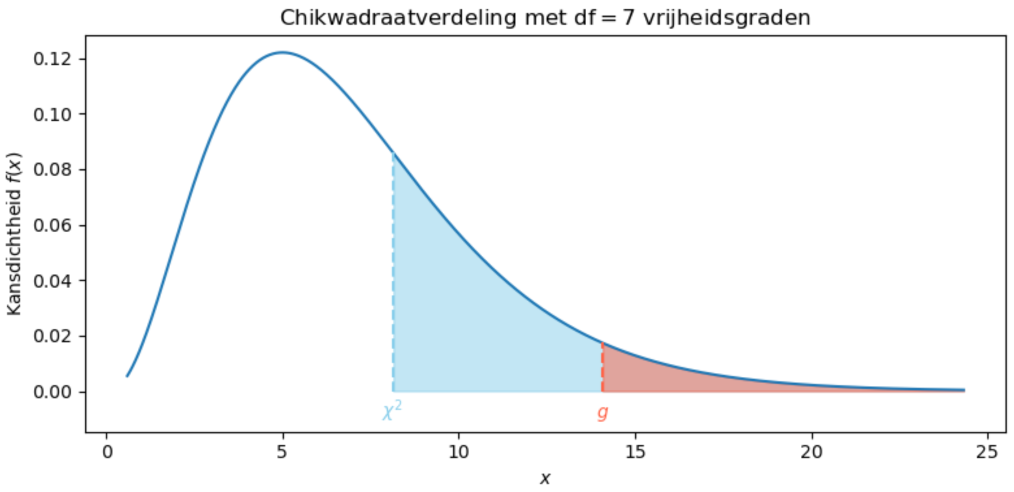


Er is onvoldoende reden om aan te nemen dat het aantal klanten in de kantine tijdens de lunchpauze een andere kansverdeling volgt dan een Poissonverdeling met $\mu = 3,7$.

Stap 4: geef een conclusie en formuleer deze in de originele context

Methode 2 (p-waarde): de **p-waarde** (rechteroverschrijdingskans) wordt gegeven door

$$p = P(X^2 \geq \chi^2) = \chi^2\text{cdf}(a = 8{,}1304; b = 10^{99}; df = 7) \approx 0{,}3212$$



→ de p-waarde is groter dan het significantieniveau $\alpha = 0{,}05$, dus H_0 wordt niet verworpen.

Er is onvoldoende reden om aan te nemen dat het aantal klanten in de kantine tijdens de lunchpauze een andere kansverdeling volgt dan een Poissonverdeling met $\mu = 3,7$.

Stap 5: formuleer een duidelijke conclusie in de originele context.

Volgens beide methoden (kritiek gebied / p-waarde) wordt er op basis van deze steekproef de nulhypothese H_0 niet verworpen.

Dat betekent dat op basis van deze steekproef onvoldoende reden is om de hypothese te verwerpen dat het aantal binnenkomende personen bij de bedrijfskantine $Poisson(\mu = 3{,}7)$ verdeeld is.

Enkele aandachtspunten

- **Vuistregel:** de chikwadraatverdeling benadert de verdeling van de toetsingsgrootheid, en is goed genoeg als alle verwachte frequenties groter of gelijk aan 5 zijn ($E_i \geq 5$). Mocht dat niet zo zijn, dan moet je rijen samenvoegen (**merk op dat dit het aantal vrijheidsgraden verlaagt!**)

Aantal personen	Verwachte frequenties E_i
0	2
1	3
2	6
3	11
4	15
5	17
6	8
≥ 7	4
Totaal	**66**

$\rightarrow$

Aantal personen	Verwachte frequenties E_i
0-1	5
2	6
3	11
4	15
5	17
≥ 6	12
Totaal	**66**

- Bij het schatten van parameters verlies je ook een vrijheidsgraad (voorbeeld: opgave 10.15)

Samenvatting

- De chikwadraatverdeling en toepassingen
  - Onafhankelijkheidstoets
  - Aanpassingstoets

Huiswerk:

- Lezen van A. Buijs: hoofdstuk 10.1 (blz. 315-318), 10.2 (blz. 318-323), 10.3 (blz. 325-330)

- Opdrachten:
  - Hoofdstuk 10: m1-m4, 10.1, 10.5a, 10.7, 10.11, 10.12, 10.13

Volgende les: verschiltoetsen