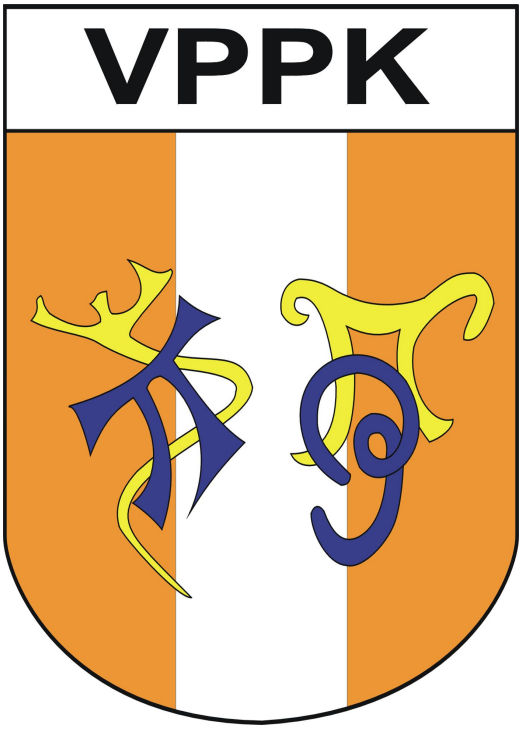
Statistiek II

Sessie 5

Feedback

Deel 5



VPPK

Universiteit Gent

2017-2018

Feedback Oefensessie 5

1 Statismex, gewicht en slaperigheid2

1. Lineair model: slaperigheid2 $= \beta_0 + \beta_1 \text{dosis} + \beta_2 \text{bd} + \varepsilon$
   $H_0 : \beta_1 = \beta_2 = 0$ vs. H_a : minstens 1 van de β's is verschillend van 0

2. Predictie voor een individu i is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{dosis} + \hat{\beta}_2 \text{bd} = 45 - 3.8\text{dosis} - 2.7\text{bd}$.
   Individu 1: $\hat{y}_1 = 45 - 3.8 \times 5 - 2.7 \times 9 = 1.7$
   Individu 2: $\hat{y}_2 = 45 - 3.8 \times 5 - 2.7 \times 7 = 7.1$
   Het verschil tussen beide is 5.4. Dit is gelijk aan $\hat{\beta}_2 \times 2$. Dit is een logisch gevolg van het feit dat we onze dosis gelijk hebben gehouden en onze bloeddruk 2 eenheden hebben laten dalen. Als we 1 eenheid bloeddruk stijgen en de rest constant houden dan dalen we $\hat{\beta}_2$ eenheden in slaperigheid2.

3. Het geobserveerde residu van een individu i is $e_i = y_i - \hat{y}_i$
   Individu 1: $e_1 = 1 - 1.7 = -0.7$
   Individu 2: $e_2 = 12 - 7.1 = 4.9$

4. $$\hat{\sigma}_\varepsilon = \frac{\sum_{i=1}^{n} e_i}{n - p - 1} = \frac{(-0.7)^2 + 4.9^2 + (-3.5^2) + (-0.1^2) + (-1.2^2) + (-1.6^2) + 1^2 + 3.4^2}{8 - 2 - 1} = \frac{53.32}{5} = 10.664$$

5. Ja. In sessie 3 hebben we al gezien dat dosis een significante predictor is voor bd. Wanneer dosis stijgt dan zien we een dalende trend in bd. De variabelen hebben een duidelijke (negatieve) correlatie. Natuurlijk is deze steekproef te klein om met zekerheid uitspraken te gaan doen.

2 Statismex, Statistine en bloeddruk

1. Afhankelijke variabele: Bloeddruk – Ratio meetniveau
   Onafhankelijke variabele: Groep – Nominaal meetniveau
   Beste techniek: ANOVA
   $H_0 : \mu_1 = \mu_2 = \mu_3$ versus H_a : minstens 1 μ is verschillend Of $H_0 : \beta_1 = \beta_2 = 0$ versus H_a : minstens 1 β is verschillend

2. We hebben $I - 1$ hulpveranderlijken nodig, waar I het aantal niveaus van de variabele is, dus $I - 1 = 3 - 1 = 2$ hulpveranderlijken. We gebruiken dummy-codering:

Bloeddruk	10	7	10	10	12	10	13	14
Groep	1	2	1	2	3	2	1	3
X_1	1	0	1	0	0	0	1	0
X_2	0	1	0	1	0	1	0	0

3. $\bar{y}_1 = \frac{10+10+13}{3} = 11 (= \hat{\mu}_1 = \widehat{\mathbb{E}(Y_{1k})})$
   $\bar{y}_2 = \frac{7+10+10}{3} = 9 (= \hat{\mu}_2 = \widehat{\mathbb{E}(Y_{2k})})$
   $\bar{y}_3 = \frac{12+14}{2} = 13 (= \hat{\mu}_3 = \widehat{\mathbb{E}(Y_{3k})})$
   Aangezien de 3de categorie het referentieniveau is:
   $\hat{\beta}_0 = \hat{\mu}_3 = 13$
   $\hat{\beta}_1 = \hat{\mu}_1 - \beta_0 = 11 - 13 = -2$
   $\hat{\beta}_2 = \hat{\mu}_2 - \beta_0 = 9 - 13 = -4$

4. Een predictie heeft de vorm $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$.

   1 $y_1 = 13 - 2 \times 1 - 4 \times 0 = 11$

   2 $y_2 = 13 - 2 \times 0 - 4 \times 1 = 9$

   3 $y_3 = 13 - 2 \times 1 - 4 \times 0 = 11$

4 $y_4 = 13 - 2 \times 0 - 4 \times 1 = 9$

5 $y_5 = 13 - 2 \times 0 - 4 \times 0 = 13$

6 $y_6 = 13 - 2 \times 0 - 4 \times 1 = 9$

7 $y_7 = 13 - 2 \times 1 - 4 \times 0 = 11$

8 $y_8 = 13 - 2 \times 0 - 4 \times 0 = 13$

We zien dus dat de predictie van een individu gelijk is aan het gemiddelde van de groep waartoe dit individu behoort.
Het residu berekenen we door $e_i = y_i - \hat{y}_i$:

1 $e_1 = 10 - 11 = -1$

2 $e_2 = 7 - 9 = -2$

3 $e_3 = 10 - 11 = -1$

4 $e_4 = 10 - 9 = 1$

5 $e_5 = 12 - 13 = -1$

6 $e_6 = 10 - 9 = 1$

7 $e_7 = 13 - 11 = 2$

8 $e_8 = 14 - 13 = 1$

5. Onder het nulmodel is de beste schatter voor $\beta_0 = \bar{y} = 10.75$. Dit is dezelfde waarde voor elk individu. De residuen worden op dezelfde manier als het voorgaande berekend:

   1 $e_1 = 10 - 10.75 = -0.75$

   2 $e_2 = 7 - 10.75 = -3.75$

   3 $e_3 = 10 - 10.75 = -0.75$

   4 $e_4 = 10 - 10.75 = -0.75$

   5 $e_5 = 12 - 10.75 = 1.25$

   6 $e_6 = 10 - 10.75 = -0.75$

   7 $e_7 = 13 - 10.75 = 2.25$

   8 $e_8 = 14 - 10.75 = 3.25$

6. Voor de F-verhouding moeten we eerst de SS_{Res} onder het lineair model en onder het nulmodel gaan berekenen:

$$SS_{\text{Res0}} = \sum_{i=1}^{n} e_i^2 = 33.5$$
$$SS_{\text{Res1}} = \sum_{i=1}^{n} e_i^2 = 14$$

   De F-verhouding onder de nulhypothese:

$$\frac{(SS_{\text{Res0}} - SS_{\text{Res1}})/((n-k-1)-(n-p-1)}{SS_{\text{Res1}}/(n-p-1)} = \frac{(33.5-14)/2}{14/5} = 3.482143$$

   In deze formule is k het aantal predictoren onder het nulmodel, in dit geval 0. p is het aantal predictoren (elke hulpveranderlijke wordt gezien als een aparte predictor) onder het lineair model, in dit geval 2.

7. De F-verdeling onder de nulhypothese heeft $k - p$ als eerste vrijheidsgraad, en $n - p - 1$ als tweede vrijheidsgraad, dus we zoeken de kans dat we onder de $F_{2;5}$-verdeling een waarde minstens even groot vinden als 3.482, dus $p = 0.1129104$

3 Gauss-Markov

1. Aangezien de vierde groep referentieniveau is en we gebruik maken van dummy-codering krijgen we dezelfde redenering als in vorige oefening. $\hat{\beta}_0$ is de waarde van het gemiddelde van het referentieniveau. In boxplots worden de medianen weergegeven, maar in dit geval gaan we er van uit dat de medianen en de gemiddeldes hetzelfde zijn - om toch de oefening te kunnen maken.

   - $\hat{\beta}_0 = 75$
   - $\hat{\beta}_1 = 71 - \hat{\beta}_0 = -4$
   - $\hat{\beta}_2 = 73 - \hat{\beta}_0 = -2$
   - $\hat{\beta}_3 = 66 - \hat{\beta}_0 = -9$

2. De 2de Gauss-Markov assumptie stelt dat de variantie niet mag afhangen van het individu (homoscedasticiteit) of van de groep. We moeten een constante variantie hebben over de groepen. We zien hier dat groep 2 een overduidelijk grotere variantie heeft dan de andere 3 groepen. Dus de 2de Gauss-Markov assumptie is geschonden.

4 Opleiding en vervoermiddel

1. Afhankelijke variabele: vervoermiddel – nominaal meetniveau
   Factor/Onafhankelijke variabele: opleiding – nominaal meetniveau

2. Variantie-analyse (of lineaire regressie met nominale variabelen) is niet geschikt, want de afhankelijke variabele moet dan minstens van intervalniveau zijn. Om dezelfde reden is een lineaire regressie en een t-test uit den boze. We kunnen hier enkel overgaan tot categorische data-analyse.

5 Tutoring

1. Naast de 3 Gauss-Markov assumpties is er ook de assumptie van normaliteit van de residuals. De distributie van de residuals is dezelfde als deze van de afhankelijke variabele. Dit komt omdat als we gaan zien naar de formule

$$Y_{ik} = \mu + \alpha_i + \varepsilon_{ik},$$

   en we hier de verwachting van nemen

$$\begin{aligned}\mathbb{E}(Y_{ik}) &= \mathbb{E}(\mu + \alpha_i + \varepsilon_{ik}) \\ &= \mathbb{E}(\mu) + \mathbb{E}(\alpha_i) + \mathbb{E}(\varepsilon_{ik}) \\ &= \mu_i,\end{aligned}$$

   zien we perfecte puntschattingen. Deze vallen telkens samen voor de respectievelijke groep. Als we de foutterm ε_{ik} terug invoegen krijgen we terug onze verdeelde observaties. Dus de verdeling van de residuals is dezelfde verdeling als degene die we zien bij de geobserveerde variabelen.
   We zien dat onze geobserveerde variabelen niet normaal verdeeld zijn in de groep, maar rechts-scheef verdeeld. Dus de assumpties zijn niet allemaal geldig.

6 Dyslexie en dyscalculie in Statisland

1. $H_0 : \hat{\pi}_{i,j} = \hat{\pi}_j$ voor alle i en j versus H_a : minstens 1 geschatte proportie verschilt tussen de populaties

2. Aangezien we nu de proporties willen vergelijken tussen populaties op 1 categorische variabele gaan we gebruik maken van de homogeniteitstest. Om de voorwaarden te testen voor meer dan 5 cellen gaan we ineens de test uitvoeren. Als we merken dat in meer dan 20% van de gevallen onze theoretische frequenties kleiner zijn dan 5 mogen we geen gebruik maken van de benadering van de χ^2.

3. De toetsingsgrootheid berekenen gebeurt in verschillende stappen. In een eerste stap bereken we de marginale frequenties op basis van de geobserveerde frequenties n_{ij}:

f_{ij}	Dyslexie	Dyscalculie	Beide	n_i
Vrouwen	40	20	10	70
Mannen	30	15	15	60
$n_{.j}$	70	35	25	130 $(= n)$

Als volgende stap gaan we nu de theoretische proporties berekenen door de marginale frequenties over geslachten heen te delen door het totale aantal observaties:

f_{ij}	Dyslexie	Dyscalculie	Beide	n_i
Vrouwen	40	20	10	70
Mannen	30	15	15	60
$n_{.j}$	70	35	25	130 $(= n)$
$\hat{\pi_{.j}}(= n_{.j}/n)$	0.5385	0.2692	0.1923	1

Nu kunnen we de theoretische frequenties berekenen door onze theoretische proporties te vermenigvuldigen met de steekproefgroottes per populatie:

f_{ij}	Dyslexie	Dyscalculie	Beide	n_i
Vrouwen	40	20	10	70
Mannen	30	15	15	60
$n_{.j}$	70	35	25	130 $(= n)$
$\hat{\pi}_{.j}(= n_{.j}/n)$	0.5385	0.2692	0.1923	1
Vrouwen $n_i\hat{\pi}_{.j}$	37.695	18.844	13.461	
Mannen $n_i\hat{\pi}_{.j}$	32.310	16.152	11.538	

We zien dat elke cel een theoretische frequentie van minstens 5 hebben, en er is dus aan de voorwaarden voldaan. Nu we alle gegevens hebben kunnen we onze toetsingsgrootheid gaan berekenen:

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(f_{ij} - n_i\hat{\pi}_{.j})^2}{n_i\hat{\pi}_{.j}} = 2.3878$$

Onder de nulhypothese is dit een realisatie van de χ^2-verdeling met $(i-1)(j-1)$ vrijheidsgraden. Dus we zoeken de kans dat we een waarde van minstens 2.3878 observeren onder de χ^2_2-verdeling, wat een $p = 0.303$ oplevert.

Extra informatie

Wanneer gebruiken we welke methode?

We gaan hier trachten een overzicht te geven over wanneer we welke methode gaan toepassen, wat de assumpties zijn die moeten voldaan zijn. Bekijk het als een beslissingsboom met een samenvatting van de volledige methode. We gaan hier niet in op details, dit is gewoon puur een overzicht.

1 variabele

Als we 1 variabele hebben moeten we ons afvragen wat deze variabele meet.

- We hebben een variabele die de uitkomst van een dichotoom proces weergeeft (bijvoorbeeld het opwerpen van een munt) ⇒ Binomiale verdeling - Binomiale test
  - Assumpties: geen
  - Methode: gebruik de formule

    $$\mathcal{P}(X \sim B(n,\pi) = k) = \frac{n!}{k!(n-k)!}\pi^k(1-\pi)^{(n-k)}$$

    Waar n staat voor het totale aantal in de steekproef (bijvoorbeeld totaal aantal worpen met een muntstuk), k staat voor de keren dat we een bepaald kenmerk observeren (bijvoorbeeld munt bij het opwerpen van een muntstuk) en π de proportie in de populatie (bijvoorbeeld 1/2)
- Onze variabele stelt een populatie voor met verschillende groepen, met in elke groep een proportie of frequentie ⇒ Pearson Chi Squared
  - Assumptie bij meer dan 5 cellen: in maximaal 20% van de gevallen mag onze theoretische frequentie ($n_i\pi_{.}j$) kleiner zijn dan 5
  - Assumptie bij minder dan of exact 5 cellen: in alle cellen moet onze theoretische frequentie ($n_i\pi_{.}j$) minstens 5 zijn.
  - Methode: we raden hier aan te volgen met de laatste oefening van deze sessie ernaast om zo de symbolen goed te begrijpen in het stappenplan
    1. We hebben een tabel met de frequenties f_{ij}
    2. We bereken de marginale frequenties, zowel over i als over j om zo tot alle $n_{i.}$ en $n_{.j}$ te komen. Tenslotte berekenen we nog n door alle $n_{i.}$ op te tellen (ter controle is dit ook de som van alle $n_{.j}$)
    3. We delen elke $n_{.j}$ door n en bekomen zo de theoretische proporties $\pi_{.j}$
    4. Nu maken we een nieuwe tabel door voor elke cel de thereotische frequenties te berekenen door de theoretische proporties te vermenigvuldigen met de steekproefgroottes per populatie ($n_{i.}\hat{\pi}_{.j}$)
    5. We nemen nu het verschil van de geobserveerde frequentie en de theoretische frequentie en kwadrateren deze. En deze delen we door de theoretische frequentie
    6. Als laatste stap tellen we de bekomen resultaten van vorige stap op

    Op deze manier hebben we de formule

    $$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(f_{ij} - n_i\hat{\pi}_{.j})^2}{n_i\hat{\pi}_{.j}}$$

    volledig doorlopen. De gevonden toetsingsgrootheid volgt een χ^2-verdeling met $(I-1)(J-1)$ vrijheidsgraden
- Onze variabele stelt een meting voor, minstens van intervalniveau ⇒ One-sample t-test

- Assumpties: ofwel moet onze variabele normaalverdeeld zijn, ofwel moet onze steekproef minstens 30 zijn (en mag onze verdeling ook niet te scheef zijn - cf. pp. 81). Bij een te grote afwijking van de normaalverdeling mogen we ook niet verder gaan. De puntschatting zal wel vrij betrouwbaar en valide zijn, maar ons betrouwbaarheidsinterval zal te groot worden (onze standaardfout wordt groter en we krijgen een inflatie van de Type I fout)
- Methode: we berekenen eerst de toetsingsgrootheid

$$T = \frac{\bar{X} - \mu_X}{S_X/\sqrt{n}}$$

Vervolgens kunnen we het $(1-\alpha)$-betrouwbaarheidsinterval berekenen:

$$\left[\bar{x} \pm t_{n-1,\alpha/2}\frac{s_X}{\sqrt{n}}\right]$$

Voor de p-waarde en meer details van het betrouwbaarheidsinterval verwijzen we naar de eerste bundel, waar we in detail hierop zijn ingegaan.

2 variabelen

Wanneer we 2 variabelen hebben, dan hebben we 5 mogelijkheden. We moeten ons nu afvragen van welk meetniveau de variabelen zijn, en wat we trachten te meten. 2 mogelijkheden komen overeen, namelijk de Welch toets en de lineaire regressie. Dit laatste hebben we in een vorige bundel al behandeld.

- 2 nominale variabelen: Dit is een uitbreiding van de Pearson Chi Squared. Dit is dezelfde werkwijze als voorheen beschreven. Het grote verschil is dat we bij 1 variabele maar 1 rij hebben. In het volgende stuk van deze bundel werken we een oefening van vorig jaar uit waar maar 1 variabele wordt gebruikt. Het geval met 2 variabelen is zoals in oefening 6 van deze bundel
- 2 variabelen van intervalniveau: Lineaire regressie
- 1 variabele van intervalniveau en 1 variabele van nominaal niveau met meer dan 2 niveaus: Lineaire regressie met nominale variabele (of ANOVA) (dit stappenplan overlopen we bij de meervoudige lineaire regressie, omdat deze hier meer gelijkenissen mee vertoont)
- 1 variabele van intervalniveau en 1 dichotome variabele: Lineaire regressie of Welch toets
- Eenzelfde variabele gemeten in 2 populaties: Welch-toets (indien onafhankelijke steekproeven) of Two sample paired t-test (indien afhankelijke steekproeven)

Noot: We weten dat het niet evident is om een concreet onderscheid te maken betreffende de t-test. We hebben hier getracht dit af te bakenen, maar een logisch inzicht is hier zeker van belang. Men kan bijvoorbeeld opperen dat we in sommige gevallen slechts over 1 variabele beschikken en deze opgedeeld wordt over de 2 populaties. Maar omdat deze doorgaans worden voorgesteld als 2 verschillende steekproeven of 2 verschillende metingen hebben we gekozen dit bij dit stuk te zetten
We gaan niet meer in op de t-test. We willen hier wel een stappenplan meegeven voor de Lineaire Regressie (zowel in de t-verdeling als in de F-verdeling). In een vorige bundel hebben we gezien en aangetoond dat deze equivalent zijn, en dus dezelfde p-waarde geven:

1. Identificeer de afhankelijke variabele $(= Y)$ en de predictor (of onafhankelijke variabele) $(= X)$
2. Bereken de predicties: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
3. Bereken alle Sums of Squares:

- $SS_X = \sum_{i=1}^{n}(x_i - \bar{x})^2$
- $SS_Y = SS_{\text{Tot}} = SS_{\text{Res0}} = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- $SS_{\text{Mod}} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- $SS_{\text{Res1}} = SS_{\text{Res}} = \sum_{j=1}^{n}(y_i - \hat{y}_i)^2$

We weten ook dat $SS_{\text{Tot}} = SS_{\text{Mod}} + SS_{\text{Res}}$

4. Bereken de toetsingsgrootheid onder de t-verdeling:

$$T = \frac{B_1}{\sqrt{\frac{SS_{\text{Res}}}{(n-2)SS_X}}}$$

Deze toetsingsgrootheid volgt een t-verdeling met $n-2$ vrijheidsgraden. De p-waarde die we berekenen op basis van de toetsingsgrootheid is hier eenzijdig, en moeten we dus altijd vermenigvuldigen met 2

5. Bereken de toetsingsgrootheid onder de F-verdeling:

$$F = \frac{SS_{\text{Res0}} - SS_{\text{Res1}}}{SS_{\text{Res1}}/(n-2)}$$

Deze volgt een F-verdeling met vrijheidsgraden $df_1 = 1$ en $df_2 = n - 2$

6. Bereken $R^2 = SS_{\text{Mod}}/SS_{\text{Tot}}$. Dit is de proportie verklaarde variantie.

Meer dan 2 variabelen

In deze cursus hebben we hier enkel de Meervoudige Lineaire Regressie gezien. Hier hebben we een afhankelijke variabele van minstens intervalniveau. Het stappenplan om dit op te lossen is zeer gelijkaardig met die van de Enkelvoudige Lineaire Regressie. De verschillen zitten in het maken van de predicties, waar we alle predictoren mee in rekening brengen. Ook is het mogelijk dat we 2 modellen met predictoren met elkaar gaan vergelijken, en dan krijgen we 2 predicties voor elk individu. Dit betekent ook dat we 2 Sums of Squares krijgen van de Residuals op basis van de predicties.
Om dit iets meer concreet te maken gaan we bij het berekenen van de toetsingsgrootheid 2 mogelijkheden geven

- Voorbeeld 1: We hebben een nulmodel (geen predictoren) en een lineair model met meerdere predictoren (wanneer we 1 nominale onafhankelijke variabele hebben gaan we elke hulpveranderlijke zien als een aparte predictor):

$$F = \frac{(SS_{\text{ResA}} - SS_{\text{ResB}})/(df_A - df_B)}{SS_{\text{ResB}}/df_B}$$

Hier is $SS_{\text{ResA}} = SS_{\text{Res0}}$. We hebben hier $df_A = n - 1$ en $df_B = n - p - 1$ waar p gelijk is aan het aantal predictoren. Deze toetsingsgrootheid volgt een F-verdeling met volgende vrijheidsgraden: $df_1 = p$ en $df_2 = n - p - 1$ met p het aantal predictoren.

- Voorbeeld 2: We hebben een model met 4 predictoren (model A) en een model met slechts 2 van de 4 predictoren (model B). We willen deze vergelijken. We krijgen dan:

$$F = \frac{(SS_{\text{ResA}} - SS_{\text{ResB}})/(df_A - df_B)}{SS_{\text{ResB}}/df_B}$$

We hebben nu $df_A = n - p - 1$, waar p staat voor het aantal predictoren in model A (dus 4). $df_B = n - k - 1$ waar k gelijk is aan het aantal predictoren in model B. Onze F-verdeling heeft nu vrijheidsgraden $df_1 = p - k$ en $df_2 = k$.

Belangrijke noot

Een F-verdeling en een χ^2 verdeling zijn beide al inherent tweezijdig. Dus we moeten de bekomen p-waarde nooit vermenigvuldigen met 2 in de contexten gezien in deze bundel. Ook is het belangrijk te weten dat we bij de F-verdeling en de χ^2-verdeling altijd gaan zoeken naar de waarde groter dan of gelijk aan ($\geq$). Dit omdat we geïnteresseerd zijn in extreme waarden, en beide verdelingen niet naar $-\infty$ gaan, maar beginnen vanaf 0.

Extra oefening

Dyslexie en dyscalculie

Tien procent van de populatie in Vlaanderen heeft dyslexie, vijf procent heeft dyscalculie en vijf procent heeft beide. Je wil de hypothese toetsen dat de verdeling dezelfde is bij Vlaamse vrouwen. Hieronder de gegevens in een steekproef van 100 vrouwen:

	Dyslexie	Dyscalculie	Beide
Frequentie	14	7	2

$P(t_3 \geq 0.1) = 0.4633, P(t_4 \geq 0.1) = 0.4626, P(t_3 \geq 4.3125) = 0.0115$
$P(\chi^2_3 \geq 4.3125) = 0.2296, P(\chi^2_4 \geq 4.3125) = 0.3654$

1. Welke toets moet je gebruiken om deze hypothese te toetsen? Is er aan de voorwaarden voldaan om deze toets te gebruiken?

2. Bereken de toetsingsgrootheid en de corresponderende p-waarde.

Oplossing

1. Aangezien we de geobserveerde frequentie in een populatie willen vergelijken met een theoretische proportie gaan we de Pearson Chi Squared toets gebruiken. Om aan de voorwaarden te voldoen moeten we, aangezien er maar 4 groepen zijn, minstens een theoretische frequentie hebben van 5 per groep. We zien

	Dyslexie	Dyscalculie	Beide	Geen
Geobserveerde frequentie (f_i)	14	7	2	77
Theoretische proportie (π_i)	0.1	0.05	0.05	0.8
Theoretische frequentie ($n\pi_i$)	10	5	5	80

   dat de theoretische frequentie voor elke categorie minstens 5 is. Dus er is aan de voorwaarden voldaan.

2. Toetsingsgrootheid:

$$\begin{aligned} X^2 &= \sum_{i=1}^{p} \frac{(f_i - n\pi_i)^2}{n\pi_i} \\ &= \frac{(14-10)^2}{10} + \frac{(7-5)^2}{5} + \frac{(2-5)^2}{5} + \frac{(77-80)^2}{80} \\ &= 4.3125 \end{aligned}$$

   Onder de nulhypothese is deze asymptotisch χ^2 verdeeld met $p-1$ vrijheidsgraden. Dus we zoeken de kans dat we een waarde van minstens 4.3125 observeren onder de χ^2_3-verdeling. We bekomen een p-waarde van $p = 0.2296$.