HOOFDSTUK VII

VARIANTIEANALYSE

1. DOELSTELLINGEN

Op het eind van dit hoofdstuk zijn studenten in staat handmatig een eenvoudige variantie analyse uit te voeren. Studenten kunnen een F-toets uitvoeren en nagaan of waargenomen verschillen al dan niet significant zijn. Studenten kunnen eta-kwadraat berekenen en correct interpreteren.

2. TE ONTHOUDEN KERNBEGRIPPEN

ANOVA (*Analysis of variance)*	Een toets voor de relatie tussen een nominale en een metrische variabele. De berekeningswijze is gebaseerd op de varianties in steekproeven.
Binnengroepsvariantie *(within-groups)*	Som van de binnengroepsvariaties delen door het aantal vrijheidsgraden (= n – aantal groepen)
Eta-kwadraat	De verhouding tussen de tussengroepsvariatie en de totale variatie in Y Equivalent voor de determinatiecoëfficiënt uit een regressie analyse % van de variatie in Y dat verklaard kan worden door X
F-ratio	Tussengroepsvariantie delen door binnengroepsvariantie Hoe groter de F-waarde, hoe groter de verschillen tussen de groepen in verhouding tot de verschillen binnen de groepen
Tussengroepsvariantie *(between-groups)*	Groepsomvang * (som van de gekwadrateerde afwijkingen van de groepsgemiddelden tov het algemene gemiddelde) delen door aantal vrijheidsgraden (=aantal groepen MIN 1)

3. STATISTISCHE SYMBOLEN EN FORMULES

F-waarde = tussengroepsvariantie / binnengroepsvariantie

$$F = \frac{\text{Verklaarde variantie}}{\text{Niet-verklaarde variantie}} = \frac{\textit{Between SS / DF}}{\textit{Within SS / DF}}$$

4. OEFENINGEN

1. In onderstaande tabel zie je het aantal geregistreerde fietsdiefstallen in 30 Belgische gemeenten in 2008, opgesplitst naar gemeentetype.
   - Is er een verband tussen het gemeentetype en het aantal fietsdiefstallen ?
   - Hoe sterk is het verband ?
   - Is dit verband significant ?
   - Hoeveel bedraagt de goodness of fit maat ?

Grootsteden	Middelgrote steden	Rurale gemeenten
3500	1850	400
2700	1650	450
2900	1450	500
3200	1600	550
3150	1550	390
3300	1800	530
2650	1400	410
4000	1750	440
3500	1250	570
3000	1500	600

Vul de Beschrijvende statistieken en ANOVA-tabel hieronder verder aan

Werkwijze

- Bereken het gemiddelde voor elke groep.
  Wat merk je reeds op ?
- Bereken de binnengroepsvariatie voor elke groep.
- Bereken de totale binnengroepsvariatie (SSwithin).
- Bepaal het aantal vrijheidsgraden. (Df_within)
- Bereken de totale binnengroepsvariantie (Mean Square within)
- Bereken de tussengroepsvariatie (SSbetween)
- Bepaal het aantal vrijheidsgraden (Df_between)
- Bereken de tussengroepsvariantie (Mean Square between)
- Bepaal de F-ratio
- Vergelijk de bekomen F-waarde met de kritieke F-waarde, gegeven het aantal vrijheidsgraden in teller en noemer en een significantieniveau α=.05.

BESCHRIJVENDE STATISTIEKEN		
GEREGISTREERDE FIETSDIEFSTALLEN		
	N	Mean
1 STADSSCHOOL		
2 GEMEENTESCHOOL		
3 PLATTELANDSSCHOOL		
Total		

ANOVA					
GEREGISTREERDE FIETSDIEFSTALLEN					
	Sum of Squares	*df*	Mean Square	*F*	Sig.
Between Groups					
Within Groups					
Total					

2. Controlevragen:

- Welke hypothesen kun je met een variantieanalyse testen ?
- Welke toetsingsgrootheid gebruik je bij de variantieanalyse ?
- Wat gebeurt er met de *F*-waarde als de tussengroepsvariantie groter is dan de binnengroepsvariantie ?
- Wat gebeurt er met de *p*-waarde naarmate de *F*-waarde groter is ? En wat betekent dit voor de conclusie ?
- Wanneer gebruik je bij regressieanalyse een *F*-toets ?
- Wat betekent het als je $p<0.05$ vindt bij deze *F*-toets ?
- Welke toets wordt nog meer bij regressieanalyse gebruikt ?

3. Twintig scholieren (tien jongens en tien meisjes) worden gemeten op een aantal variabelen zoals duur van afwezigheid gedurende afgelopen jaar en testscores bij het begin van het schooljaar. Je vindt hieronder de gegevens.

Biologisch geslacht Code 1= jongen Code 2 = meisje	Duur afwezigheid (uitgedrukt in aantal dagen)	Testscores
1	24	13
1	20	16
1	8	7
1	12	30
1	5	5
1	24	10
1	0	9
1	8	15
1	20	18
1	24	20
2	7	18
2	30	14
2	2	20
2	10	9
2	18	13
2	9	13
2	20	10
2	10	16
2	15	5
2	8	7

Voer een variantieanalyse uit tussen jongens en meisjes op 'duur afwezigheid' en op 'testscores'.

- Bepaal de *F*-ratio voor beide variantieanalyses
- Vergelijk de bekomen *F*-waarde met de kritieke *F*-waarde, gegeven het aantal vrijheidsgraden in teller en noemer en een significantieniveau α=.05.
- In geval van een significant resultaat, bepaal de waarde van de goodness of fit maat.
- Interpreteer de resultaten.

BESCHRIJVENDE STATISTIEKEN		
DUUR AFWEZIGHEID		
	N	Mean
1 JONGENS		
2 MEISJES		
Total		

ANOVA					
DUUR AFWEZIGHEID					
	Sum of Squares	*df*	Mean Square	*F*	Sig.
Between Groups					
Within Groups					
Total					

BESCHRIJVENDE STATISTIEKEN		
TESTSCORES		
	N	Mean
1 JONGENS		
2 MEISJES		
Total		

ANOVA					
TESTSCORES					
	Sum of Squares	*df*	Mean Square	*F*	Sig.
Between Groups					
Within Groups					
Total					

4. Zesendertig personen namen deel aan een experiment om de effecten van alcohol op het rijvermogen te ontdekken. Ze werden willekeurig toegewezen aan drie verschillende condities: placebo (geen alcohol), lage alcohol en hoge alcohol. Het niet-alcoholische drankje zag er precies hetzelfde uit en smaakte hetzelfde als de andere drankjes(!). Deelnemers werden gewogen en kregen de passende hoeveelheid drank. Na een half uur drinken reden de deelnemers tien minuten in een simulator, en het aantal gemaakte fouten werd automatisch geregistreerd door de computer. De gegevens staan vermeld in onderstaande tabel.

   H0 = Er is geen verband tussen alcohol en rijvermogen.
   Ha = Er is een verband tussen alcohol en rijvermogen.

Observaties	Placebo	Lage alcohol	Hoge alcohol
1	5	5	8
2	10	7	10
3	7	9	8
4	3	8	9
5	5	2	11
6	7	5	15
7	11	6	7
8	2	6	11
9	3	4	8
10	5	4	8
11	6	8	17
12	6	10	11

- Voer een variantie-analyse uit.
- Kunnen we de nulhypothese verwerpen dat er geen verband is tussen leeftijd en delict-type?

ANOVA

RIJVERMOGEN

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups					
Within Groups					
Total					

- Hoeveel bedraagt de goodness of fit maat? Hoe sterk is het verband?

5. In onderstaande tabel vind je de leeftijd van 30 veroordeelde daders van witteboordencriminaliteit. De veroordeelden werden ingedeeld volgens delict-type: fraude – omkoping – witwassen. We willen nagaan of de drie groepen statistisch significant verschillen naar gemiddelde leeftijd. Met andere woorden: is er een verband tussen leeftijd en delict-type?

H0 = Veroordeelde daders van fraude, omkoping en witwassen verschillen niet statistisch significant naar gemiddelde leeftijd (er is geen verband tussen leeftijd en delict-type)

Ha = Veroordeelde daders van fraude, omkoping en witwassen verschillen statistisch significant naar gemiddelde leeftijd (er is een verband tussen leeftijd en delicttype)

Observaties	FRAUDE	OMKOPING	WITWASSEN
1	19	28	35
2	21	29	46
3	23	32	48
4	25	40	53
5	29	42	58
6	30	48	61
7	31	58	62
8	35	58	62
9	42	64	62
10	49	68	75
N	10	10	10

- Voer een variantie-analyse uit.
- Kunnen we de nulhypothese verwerpen dat er geen verband is tussen leeftijd en delict-type?

ANOVA

LEEFTIJD

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups					
Within Groups					
Total					

- Hoeveel bedraagt de goodness of fit maat? Hoe sterk is het verband?

6. Veronderstel volgende fictieve scenario:
   Een groep laatstejaarsstudenten besluit om hallucinogene drugs te nemen tijdens de colleges. Aan het eind van het semester is er een examen.
   De studenten die drugs gebruiken tijdens de colleges behaalden volgende cijfers (%):
   23 89 62 11 76 28 45 52 71 28

   De studenten die GEEN drugs gebruiken tijdens de colleges behaalden volgende cijfers (%):
   45 52 68 74 55 62 58 49 42 57

   - Wat is de onafhankelijke variabele in dit fictieve scenario?
   - Wat is de afhankelijke variabele in dit fictieve scenario?
   - Welke groep heeft de hoogste gemiddelde score?
   - In welke groep is de variatie het grootst?
   - Is er een statistisch significant verschil tussen beide groepen voor wat de gemiddelde score betreft?
   - Als er een statistisch significant verschil is, hoe groot is het effect?

7. *Heeft blootstelling aan anti-pest interventiecampagnes een impact op de kennis van jongeren over pesten op school?*

In een experimenteel onderzoek wordt nagegaan of blootstelling aan verschillende anti-pest interventiecampagnes een impact heeft op kennis van jongeren over pesten op school.

De experimentele setting omvat drie groepen respondenten:

Groep 1 wordt blootgesteld aan anti-pest campagne nummer 1.

Groep 2 wordt blootgesteld aan anti-pest campagne nummer 2.

Groep 3 wordt niet blootgesteld aan een anti-pest campagne en is de controlegroep.

Elke groep telt 9 onderzoekseenheden. In de tabel hieronder vind je de score van elke respondent in elke groep op een anti-pest kennistest afgenomen na blootstelling aan de anti-pest campagne.

ONDERZOEKSVRAAG

Verschillen groepen die verschillend zijn *blootgesteld aan anti-pest-campagnes* significant van elkaar op gemiddelde *scores op een anti-pest-kennistest*?

	GROEP 1 Campagne 1	GROEP 2 Campagne 2	GROEP 3 controlegroep
1	13	12	8
2	4	14	5
3	7	14	12
4	14	15	6
5	16	19	11
6	12	23	10
7	16	14	17
8	13	21	7
9	13	16	11

OPDRACHT

Voer een variantie analyse uit en beantwoord onderstaande vragen.

ANOVA

Scores op anti-pest-kennistest

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups					
Within Groups					
Total					

- Zijn de verschillen tussen de groepen groter dan de verschillen binnen de groepen?

- Hoeveel bedraagt de *F*-statistiek?

- Hoeveel bedraagt de kritieke *F*-waarde, gegeven aantal *DF* in teller en noemer?

- Kan de nulhypothese verworpen worden dat er *geen* verschillen zijn in gemiddelde scores op een anti-pest-kennistest tussen de groepen?

- Hoeveel bedraagt Eta-squared?

- Wat betekent dit concreet? Rapporteer in eigen woorden.

- Hoe sterk is het verband tussen blootstelling aan anti-pest interventiecampagnes en scores op een anti-pest-kennistest?

  a. Welke statistiek bereken je?

  b. Interpreteer.

8. WAAR OF VALS

UITSPRAAK	WAAR	VALS
Associatiematen hebben in geval van nominale variabelen geen richting.		
In de samenhang tussen religie en aantal jaren opleiding is de richting van de samenhang niet van toepassing.		
Een toename in opleiding hangt samen met een toename in salaris. Dit is een voorbeeld van een negatieve samenhang.		
Phi kan gebruikt worden om de samenhang te bestuderen tussen etniciteit en politieke voorkeur.		
Eta² wordt gebruikt wanneer de onafhankelijke variabele categorisch is en de afhankelijke ook.		
ANOVA kan enkel toegepast worden ingeval twee groepsgemiddeldes worden vergeleken.		
De binnengroepsvariatie is de totale hoeveelheid variatie van alle eenheden in een steekproef in verhouding tot de subgroepgemiddelden.		
De binnengroepsvariatie is de hoeveelheid variatie van alle eenheden in een steekproef in verhouding tot het algemene gemiddelde.		
Eta² wordt berekend wanneer de onafhankelijke variabele van het metrische meetniveau is en de afhankelijke categorisch.		