

1. Hieronder staan de uitkomsten van een onderzoek naar het aantal nevenfuncties van parlementariërs. Wat is de modus van deze observaties?

0, 1, 1, 1, 1, 2, 3, 4, 5, 7, 8

a. 1

b. 2

c. 3

d. de modus is in dit geval niet gedefinieerd

2. Bij variabelen die scheef verdeeld zijn, wordt vaak NIET het gemiddelde als maat voor locatie gebruikt, maar een alternatief (b.v. de mediaan of het getrimde gemiddelde). Wat is een reden om in deze gevallen NIET voor het gemiddelde te kiezen

a. Er is een directe relatie tussen het gemiddelde en het totaal (de som van de uitkomsten).

b. Het gemiddelde kan in dit soort gevallen negatief worden.

c. Bij scheve verdelingen is de afstand tussen de observaties en het gemiddelde te groot.

d. Het gemiddelde geeft bij scheve verdelingen geen goed beeld van de locatie van de observaties omdat het teveel beïnvloed wordt door de waarde van de extreme observaties.

3. Indien met toetst met een significantieniveau (α) van 0.10 betekent dit dat de kans om de nulhypothese te verwerpen wanneer deze waar is gelijk is aan 0.10. Stel dat een onderzoeker 10 toetsen uitvoert telkens met als significantieniveau 0.10 en stel dat bij al deze toetsen de nulhypothese waar is. Wat is de kans dat hij minstens 2 significante resultaten vindt?

a. 0.194

b. 0.200

c. 0.264

d. 0.736

4. Stel dat de onderzoeker werkt in een onderzoeksveld waarin bekend is dat 80% van de nulhypotheses die onderzocht worden waar is (even voor de goede orde zo'n onderzoeksveld bestaat niet; het is nooit precies bekend hoeveel nulhypotheses correct zijn, maar er zijn wel situaties waar je op theoretische gronden hier iets over kunt aannemen...). Slechts in 20% van de gevallen is de alternatieve hypothese correct. De onderzoeker besluit om een willekeurig gekozen nulhypothese te toetsen. Hij kiest als significantieniveau (α) 0.10. De toets die hij gebruikt heeft een power van 0.90, wat

inhoudt dat de kans dat hij de nulhypothese verwierpt wanneer deze niet waar is (maar dus de alternatieve hypothese waar is) gelijk is aan 0.90. De uitkomst van zijn toets is dat hij de nulhypothese moet verwierpen. Wat is de kans dat in dit geval de alternatieve hypothese correct is en het verwierpen van de nulhypothese dus terecht is?

- a. 0.200
- b. 0.308
- c. 0.692
- d. 0.900

5. Een bepaalde variabele (neem bijvoorbeeld de gemiddelde hoeveelheid kleingeld die men gedurende een week in de portemonnee heeft) is normaal verdeeld met als gemiddelde 6 en als standaarddeviatie 2. Wat is de kans dat een willekeurig gekozen observatie groter is dan 5?

- a. 0.309
- b. 0.500
- c. 0.691
- d. 0.900

6. Stel dat er twee onderzoekers een studie verrichten naar de bij **5.** genoemde variabele. Onderzoeker A neemt een willekeurig gekozen steekproef uit de gehele populatie. Onderzoeker B besluit een steekproef te nemen uit een homogener subpopulatie (een onderdeel van de populatie waar de te verwachten verschillen kleiner zijn) waarvan bekend is dat het gemiddelde van deze subpopulatie gelijk is aan het gemiddelde van de gehele populatie. De gemiddeldes in de populatie van onderzoeker A en onderzoeker B zijn dus gelijk maar de standaarddeviatie is 2 in de populatie van onderzoeker A en de standaarddeviatie is 1 in de populatie van onderzoeker B. Als onderzoeker A een steekproef kiest met een omvang van 16, hoe groot moet de steekproef van onderzoeker B dan zijn om een even betrouwbare schatting van het gemiddelde te krijgen?

- a. 4
- b. 8
- c. 16
- d. 32

Een onderzoekster doet onderzoek naar de hype rond milieuneutraal, CO₂ neutraal, klimaatneutraal en alle andere opties die het bedrijfsleven tegenwoordig biedt om als consument op een makkelijke manier wat aan een duurzame samenleving te doen. Je kunt

tegenwoordig milieuneutraal vliegen, milieuneutraal autorijden, milieuneutrale kleding kopen etc. Aangezien Eindhoven Airport besloten heeft om alle reizigers de mogelijkheid te geven om middels een betaling met een creditcard bij vertrek een compensatie te betalen voor een klimaatneutrale vlucht beperkt ze haar onderzoek tot het vliegverkeer. Wanneer iemand besluit om een compensatie te betalen dan wordt dat geld overgemaakt naar een speciaal bedrijf (GreenSeat) dat dit geld besteed aan speciale projecten die de uitgestoten CO₂ compenseert door bijvoorbeeld de aanplant van bomen.

Ze besluit om na te gaan waarom personen kiezen voor een klimaatneutrale vlucht, met andere woorden waarom ze bereid zijn om deze compensatie te betalen. Haar hypothese is dat mensen bij kortere vluchten sneller bereid zullen zijn om deze compensatie te betalen omdat er voor zo'n korte vlucht voldoende minder vervuilende alternatieven beschikbaar zijn. De reden voor het vliegtuig voor zo'n korte vlucht is dan vooral comfort en dat argument zou zo zwak zijn dat personen bij zo'n korte vlucht graag hun geweten afkopen. Bij langere vluchten is er geen alternatief en dus ook geen reden om een compensatie te betalen.

7. Als een eerste toets van haar hypothese besluit de onderzoekster om 100 willekeurig gekozen passagiers vertrekkend van Eindhoven Airport te vragen waarnaar ze vertrekken en of ze gekozen hebben voor de betaling van de compensatie voor de CO₂ uitstoot. Ze deelt de vluchten in naar lange vluchten (meestal gaat het om retours: totale vluchtlengte > 1000 kilometer) en korte vluchten (totale vluchtlengte ≤ 1000 kilometer) en ze wil weten over er verschil zit tussen deze twee vluchttypen in het relatieve aantal personen dat kiest voor compensatie. Wat is hiervoor de meest geëigende toets? NB ga hierbij uit van de door de onderzoekster gekozen opzet!!

- a. een toets op de correlatiecoëfficiënt
- b. een toets voor twee proporties
- c. een t-toets voor twee onafhankelijke steekproeven
- d. een (non-parametrische) Wilcoxon toets

8. Ze besluit om de data ook op een andere manier te analyseren. Ze berekent op basis van de vluchtlengte voor alle personen de hoogte van het compensatiebedrag. Dit noemt

ze de hypothetische compensatie (afgekort hypcomp). Vervolgens berekent ze voor de groep die wel gecompenseerd heeft (aangegeven met WEL) het gemiddelde (hypothetische) compensatie en hetzelfde doet ze voor de groep die besloten heeft om niet te compenseren (aangegeven met NIET). Als het gemiddelde hypothetische compensatiebedrag in de groep WEL groter is dan in de groep NIET betekent dit dat langere reizen relatief vaker gecompenseerd worden. Immers een hogere hypothetische compensatie betekent een langere reis en de betaalde compensaties (= hypothetische compensaties in de groep WEL) zijn dan groter dan de niet gecompenseerde bedragen (= hypothetische compensaties in de groep NIET). De uitkomsten zijn als volgt:

Voor de groep WEL: $n = 8$, gemiddelde hypothetische compensatie ($\overline{\text{hypcomp}}$): €20,-, standaarddeviatie (s_{hypcomp}): €5,-.

Voor de groep NIET: $n = 8$, $\overline{\text{hypcomp}}$: €15,-, s_{hypcomp} : €5,-.

Bovendien berekende ze verschillen (aangegeven met “verschil”). De uitkomsten hiervan waren: gemiddelde verschil: €5,-; standaarddeviatie €2,-

Toets de nulhypothese “er zit geen verschil in de populatiegemiddeldes van het hypothetische compensatiebedrag voor de populatie compenserende passagiers en de populatie niet-compenserende passagiers”. Wat is de uitkomst?

- De toetsingsgrootte t is gelijk aan 2 met 14 vrijheidsgraden. Er is geen significant resultaat en we hebben dan ook geen verschil kunnen aantonen.
- De toetsingsgrootte z is gelijk aan 2. Er is een significant resultaat en we hebben dan ook verschil kunnen aantonen.
- De toetsingsgrootte t is gelijk aan 0.884 met 14 vrijheidsgraden. Er is geen significant resultaat en we hebben dan ook geen verschil kunnen aantonen.
- De toetsingsgrootte t is gelijk aan 0.884 met 7 vrijheidsgraden. Er is geen significant resultaat en we hebben dan ook geen verschil kunnen aantonen.

9. Welke uitspraak is waar? Indien men verschillende t -toetsen voor twee onafhankelijke steekproeven die berekend zijn in verschillende onderzoeken (met andere steekproefomvang, gemiddeldes, standaarddeviaties) met elkaar vergelijkt dan geldt **ALTIJD** (ongeacht de steekproefomvang, standaarddeviaties etc.):

- a. Hoe kleiner de p-waarde, hoe groter het verschil tussen de twee steekproefgemiddeldes.
- b. Hoe kleiner de p-waarde, hoe verder betrouwbaarheidsinterval voor het verschil van de populatiegemiddeldes van 0 af ligt.
- c. Hoe kleiner p-waarde, hoe onwaarschijnlijker de nulhypothese is.
- d. Hoe kleiner de p-waarde, hoe onwaarschijnlijker de gevonden resultaten zijn ervan uitgaande dat de nulhypothese waar is.

10. Bij een andere deelstudie vroeg ze mensen die in de afgelopen periode zowel een lange als een korte vlucht hadden gemaakt of ze, en zo ja bij welke, bij hun vluchten gecompenseerd hadden. Aangezien personen die zowel de korte als de lange vlucht gecompenseerd hadden, geen informatie geven over de hypothese dat personen eerder een korte dan een lange vlucht zullen compenseren, heeft ze deze buiten beschouwing gelaten (dit is een valide procedure). Er resteerden 10 personen die zowel een lange als een korte vlucht gemaakt hadden en maar één van de twee gecompenseerd hadden. Van deze 10 had 1 persoon niet gecompenseerd op korte vlucht, en de resterende 9 hadden niet gecompenseerd op lange vlucht. Hoe moet dit resultaat geïnterpreteerd worden in het licht van de hypothese die de onderzoekster geformuleerd heeft (zie nogmaals het stuk voor opgave 7.)

- a. De p-waarde is kleiner dan 0.05; dit resultaat is bewijs voor haar hypothese.
- b. De p-waarde is kleiner dan 0.05; dit resultaat is bewijs tegen haar hypothese.
- c. De p-waarde is groter dan 0.05; dit resultaat is bewijs tegen haar hypothese.
- d. De p-waarde is groter dan 0.05; dit resultaat is niet te interpreteren.

11. De onderzoekster besluit het onderzoek van vraag 8. te herhalen bij een grotere steekproef. Ze vraagt nu aan 50 willekeurig gekozen passagiers de reisbestemming en of men gecompenseerd heeft (deze laatste variabele wordt aangegeven met “compensatie”). Ze berekent wederom de hypothetische compensatie (hypcomp). Vervolgens voert ze met SPSS een t-toets. Delen van de output volgen:

Group Statistics

compensatie		N	Mean	Std. Deviation	Std. Error Mean
hypcomp	NIET	30	35.5654	12.83272	2.34292
	WEL	20	33.2089	12.13194	2.71278

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
hypcomp	Equal variances assumed	.114	.737	.650	48	.519	2.35654	3.62576	-4.93354	9.64662
	Equal variances not assumed			.657	42.444	.514	2.35654	3.58448	-4.87499	9.58807

De onderzoekster toetst nogmaals als nulhypothese “er zit geen verschil in de populatiegemiddeldes van het hypothetische compensatiebedrag voor de populatie compenserende passagiers en de populatie niet-compenserende passagiers”.

Welke van onderstaande uitspraken is juist?

- De nulhypothese kan niet verworpen worden want de p-waarde is 0.737.
- De nulhypothese kan wel verworpen worden want de p-waarde is 0.737.
- De nulhypothese kan niet verworpen worden want de p-waarde is 0.519.
- De nulhypothese kan wel verworpen worden want de p-waarde is 0.519.

12. Stel ze zoekt in deze grotere steekproef naar personen die recentelijk (minstens) twee vluchten gemaakt hebben waarvan ze er (minstens) één wel en (minstens) één niet gecompenseerd hebben. Van deze personen vraagt ze de reisbestemming van de (meest recente) vlucht die ze wel en de (meest recente) vlucht die ze niet gecompenseerd hebben. Ze berekent zo per persoon een hypothetisch compensatiebedrag voor een wel gecompenseerde vlucht en een niet gecompenseerde vlucht. Ze wil deze gegevens weer gebruiken om de nulhypothese “er zit geen verschil in de populatiegemiddeldes van het hypothetische compensatiebedrag voor de populatie compenserende passagiers en de populatie niet-compenserende passagiers” te toetsen. Wat is de meest geëigende toets?

- a. een toets op de correlatiecoëfficiënt
- b. een toets voor twee proporties
- c. een t-toets voor twee onafhankelijke steekproeven
- d. een gepaarde t-toets

13. Wanneer je als onderzoeker data verzameld hebt en je vindt bijvoorbeeld dat het gemiddelde in de ene steekproef duidelijk groter is dan het gemiddelde in de andere steekproef, is het verleidelijk om geen ongerichte hypothesen te toetsen (H_0 : “de populatiegemiddeldes zijn gelijk aan elkaar” versus H_1 : “de populatiegemiddeldes zijn niet gelijk aan elkaar”), maar gerichte hypothesen (H_0 : “het populatiegemiddelde van populatie A is gelijk aan dat van populatie B, of in ieder geval niet groter” versus H_1 : “het populatiegemiddelde van populatie A is groter dan dat van B”). Waarom is dit NIET toegestaan?

- a. Als je dit doet is kans op type I fout te klein.
- b. Als je dit doet is kans op type I fout te groot.
- c. Als je dit doet is kans op type II fout te klein.
- d. Als je dit doet is kans op type II fout te groot.

14. De onderzoekster berekende de hypothetische compensatie door het gebruik van een formule. Ze berekende eerst de reisafstand en deze zette ze via een lineaire formule om in een hoeveelheid uitgestoten CO_2 . (per 1000 km vliegen 0.18 ton CO_2). Dit getal wordt vervolgens verdubbeld om te compenseren voor de overige uitstoot. Daarna wordt de hoeveelheid CO_2 omgezet naar euro's: €13,- per ton CO_2 . Bij dit bedrag wordt vervolgens €1.50 opgeteld voor administratiekosten. (NB in het “echt” gaat het iets ingewikkelder: er zijn verschillende uitstoot formules voor lange en korte vluchten en bij het bedrag komt ook nog BTW etc. Voor deze opgave gaan we uit van de hiervoor beschreven procedure!). Stel ze zou besluiten om een regressieanalyse uit te voeren met als afhankelijke variabele het hypothetische compensatiebedrag en als onafhankelijke variabele de reisafstand. Wat zal ze dan NIET vinden?

- a. Een (multiple) correlatie van 1.
- b. Een intercept van 0
- c. Een residuele variantie van 0.
- d. Een positief regressiegewicht voor de variabele reisafstand.

Een mogelijk storende effect kunnen inkomensverschillen zijn. Het inkomen kan zowel de vluchtbestemming (en daarmee de reisafstand en dus de hypothetische compensatie) beïnvloeden alsmede de geneigdheid om te compenseren (meerverdienende mensen kunnen sneller/minder snel geneigd zijn om te compenseren). Vandaar dat ze besluit om inkomen ook in het onderzoek te betrekken.

15. Eerst gaat ze de relatie tussen het inkomen en het hypothetische compensatiebedrag na. De correlatie in de steekproef (van 50 personen) bedraagt 0.8. Wat is het 95% betrouwbaarheidsinterval voor de populatiecorrelatiecoëfficiënt?

- a. [0.65; 0.89]
- b. [-0.2; 0.95]
- c. 0.8 ± 0.1
- d. dat is met deze gegevens niet te berekenen

Vervolgens voert ze een regressieanalyse uit waarbij ze drie blokken opneemt. De afhankelijke variabele is telkens het hypothetische compensatiebedrag (hypcomp). In het eerste blok zit inkomen (jaarinkomen in euro's) als onafhankelijke variabele, in het tweede blok komt daar compensatie bij (0 = NIET gecompenseerd, 1 = WEL gecompenseerd) bij en in het derde blok wordt tenslotte het product van inkomen en compensatie toegevoegd; deze laatste variabele heet 'interact'. De tabel met de coëfficiënten staat hieronder:

Coefficients ^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10.644	1.068		9.966	.000
	inkomen	.000	.000	.881	25.849	.000
2	(Constant)	11.367	.774		14.688	.000
	inkomen	.000	.000	1.007	38.158	.000
	compensatie	-9.529	.715	-.352	-13.319	.000
3	(Constant)	11.208	.969		11.566	.000
	inkomen	.000	.000	1.014	28.219	.000
	compensatie	-9.104	1.708	-.336	-5.331	.000
	interact	-7E-006	.000	-.020	-.274	.784

a. Dependent Variable: hypcomp

16. Het regressiegewicht voor inkomen is in alle modellen gelijk aan 0.000 (volgens de tabel). Is dit een reden om inkomen niet in het regressiemodel op te nemen?

- Ja, een regressiegewicht van 0 betekent dat de variabele geen invloed heeft.
- Ja, want de standaarderror is ook gelijk aan 0.000.
- Nee, want de waarde van het regressiegewicht voor de interactieterm is niet 0 maar -7E006.
- Nee, de grootte van het regressiegewicht heeft te maken met de schaal waarop de variabele inkomen te maken heeft.

17. Welke conclusie is juist?

- Op basis van deze analyse kan gesteld worden dat het gemiddelde gecompenseerde bedrag significant hoger is dan het gemiddelde niet-gecompenseerde bedrag en dus langere reizen vaker gecompenseerd worden dan korte reizen.
- Op basis van deze analyse kan gesteld worden dat het gemiddelde gecompenseerde bedrag significant lager is dan het gemiddelde niet-gecompenseerde bedrag en dus korte reizen vaker gecompenseerd worden dan lange reizen.
- Op basis van deze analyse kan gesteld worden dat er geen significant verschil zit tussen het gemiddelde gecompenseerde bedrag en het gemiddelde niet-gecompenseerde bedrag en er dus niet gesteld kan worden dat langere reizen vaker gecompenseerd worden dan korte reizen of het omgekeerde.
- Vanwege de interactie kan er geen uitspraken gedaan worden over welk gemiddelde hoger is.

18. Als men wil nagaan of aan de belangrijkste assumpties van het regressiemodel is voldaan dan kon met het beste kijken naar:

- a. de ANOVA tabel
- b. de tabel met de coëfficiënten
- c. de multiple correlatiecoëfficiënt
- d. de residuen

19. Wat zou er structureel met deze schatting van het regressiegewicht voor compensatie gebeuren wanneer de onderzoekster een tweemaal zo grote steekproef had genomen?

- a. Deze zou structureel groter zijn geworden, want hoe groter de steekproef hoe groter het regressiegewicht.
- b. Deze zou structureel kleiner zijn geworden, want hoe groter de steekproef hoe kleiner het regressiegewicht.
- c. Deze zou niet structureel veranderen want het regressiegewicht is een goede schatter (een schatter zonder bias) voor het populatie regressiegewicht ongeacht de steekproefomvang.
- d. Dat is niet met zekerheid te zeggen.

CASE

In de onderstaande output worden de met SPSS berekende resultaten van de analyse van de gegevens van een onderzoek gepresenteerd. In dit onderzoek werd een groep rokers die een stoppoging ondernamen gevolgd. Als afhankelijke variabele wordt telkens gekeken naar de tijd die verstrijkt vanaf het begin van de stoppoging tot aan de eerste dag van een onafgebroken periode van een week waarin de (ex-)roker geen ontwenningsverschijnselen meer ervaart; deze tijd wordt 'stoptijd' genoemd. Stel een roker stopt op 1 februari en heeft in de periode van 19 tot en met 26 februari geen ontwenningsverschijnselen. Dan is de waarde van de afhankelijke variabele stoptijd 19. Bij personen die nicotinepleisters plakken wordt als eerste dag van de stoppoging de eerste dag ZONDER pleisters genomen; de dagen dat ze een pleister plakken krijgen ze immers nog nicotine binnen!

Er nemen 40 personen aan het onderzoek deel. De onderzoekers besloten dat het niet realistisch was om mensen random aan het al dan niet plakken van nicotinepleister toe te wijzen. Bovendien wilden ze geen groep in hun onderzoek betrekken die pleisters zonder nicotine plakten; dit komt in de praktijk immers niet voor en ze wilden een realistische schatting van de stoptijd zoals die in de praktijk ook voor zal komen maken. Daarom besloten ze om de rokers zelf te laten kiezen of ze al dan niet een pleister wilden plakken. Er waren 20 rokers die er voor kozen om een pleister te plakken en 20 die dat niet wilden doen.

Even wat methodologische en statistische kanttekeningen:

- het is maar de vraag of dit de meest logische opzet is (methodologisch gezien), maar daar hoeft je in dit tentamen niet mee bezig te houden!
- we gaan ervan uit dat er niemand uitvalt, d.w.z. dat er niemand begint te roken voor ze de week zonder ontwenningverschijnselen hebben gehaald
- er bestaan betere analysetechnieken (survivalanalyse) dan we hier gebruiken, alhoewel de hier gepresenteerde analyse niet verkeerd is.

Deze kanttekeningen kun je wat mij betreft ook negeren!

Eerst werd een t-toets uitgevoerd om te kijken of er een verschil in stoptijd tussen de pleisterplakkers en de niet-pleisterplakkers is:

Group Statistics

	PLEISTER	N	Mean	Std. Deviation	Std. Error Mean
tijd tot eerste week zonder ontwenningverschijnselen	nee	20	16.70	3.706	.829
	ja	20	23.40	5.079	1.136

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
tijd tot eerste week zonder ontwenningverschijnselen	Equal variances assumed	4.727	.036	-4.767	38	.000	-6.70	1.406	-9.548	-3.856
	Equal variances not assumed			-4.767	34.762	.000	-6.70	1.406	-9.557	-3.847

a. (4 punten) Wat is op basis van deze t-toets de conclusie over het plakken van nicotinepleisters? Helpen die de stoptijd te verkorten? Gebruik $\alpha = 0.05$.

Omdat de groepen niet random zijn samengesteld kunnen er tussen de groepen verschillen in bepaalde variabelen zijn die voor een versturende werking zorgen. Er werden meerdere variabelen onderzocht, zoals geslacht, leeftijd, hoeveelheid sigaretten die per dag gerookt werd etc. Voor al deze variabelen bleek dat ze niet verschilden tussen de groep pleisterplakkers en de groep die geen pleisters plakte. De enige variabele die wel verschillend was tussen de twee groepen was het aantal jaren dat men gerookt had:

Group Statistics

	PLEISTER	N	Mean	Std. Deviation	Std. Error Mean
aantal jaren gerookt	nee	20	16.22	3.596	.804
	ja	20	25.22	4.791	1.071

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
aantal jaren gerookt	Equal variances assumed	2.795	.103	-6.720	38	.000	-9.00	1.339	-11.7	-6.289
	Equal variances not assumed			-6.720	35.25	.000	-9.00	1.339	-11.7	-6.282

Bovendien bleek het aantal jaren dat iemand gerookt had positief samen te hangen met de stoptijd; de correlatie was erg hoog (0.90) en duidelijk significant.

b. (4 punten) Leg uit waarom een variabele pas een versturende variabele kan zijn wanneer deze variabele zowel een verschil vertoont in de groep pleisterplakkers en niet-plakkers, als samenhangt met stoptijd.

Vervolgens werd onderzocht of er gecorrigeerd kon worden voor de verschillen in aantal jaren dat de persoon gerookt had. Dat corrigeren gebeurt met een regressiemodel. De variabele PLEISTER is een dummyvariabele met de waarde '0' indien de roker geen

pleister plakte en de waarde '1' indien hij/zij dat wel deed. Verder is de variabele STOP de stoptijd en de variabele ROKER staat voor het aantal jaren dat men gerookt had. Tenslotte werd een interactievariabele gedefinieerd; deze variabele heet INTERACT en is gelijk aan het product van PLEISTER en ROKER.

Men begon met een regressiemodel met als predictoren PLEISTER, ROKER en INTERACT.

De tabel met coëfficiënten zag er als volgt uit.

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.526	2.021		.755	.455
	PLEISTER	-2.427	3.096	-.222	-.784	.438
	ROKER	.935	.122	1.043	7.681	.000
	INTERACT	.028	.152	.067	.185	.855

a. Dependent Variable: STOP

c (4 punten) Hoe kan op basis van deze tabel tot de conclusie komen dat het gerechtvaardigd is om een regressiemodel te gebruiken om te corrigeren voor het aantal jaren dat met gerookt heeft?

Vervolgens werd een model gebruikt met als predictoren ROKER en PLEISTER. De tabel met coëfficiënten zag er als volgt uit:

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.234	1.244		.992	.328
	PLEISTER	-1.880	.881	-.172	-2.133	.040
	ROKER	.953	.072	1.063	13.214	.000

a. Dependent Variable: STOP

d. (4 punten) Wat is de verandering in stoptijd wanneer een persoon 1 jaar langer gerookt heeft/zou hebben?

e. (4 punten) Wat is op basis van deze regressieanalyse de conclusie over het plakken van nicotinepleisters? Vergelijk je conclusie met de conclusie van a. Gebruik $\alpha = 0.05$.

f. (4 punten) Kun je met behulp van een schets van de scatterplot van stoptijd tegen aantal jaren dat men gerookt heeft, uitleggen hoe de correctie in zijn werk gaat?

1. A; de modus is de meest voorkomende waarde.
2. D; A is een reden om het gemiddelde wel te gebruiken, B is onzin en C is niet juist: uitgedrukt in de afstand tot aan de maat voor locatie wint het gemiddelde altijd)
3. C; zie tabel IIIc uit het boek
4. C; gebruik de regel van Bayes.
5. C; De z-waarde is -0.5 $P(Z > -0.5) = P(Z < 0.5) = 1 - P(Z > 0.5)$
6. A; een even betrouwbare schatting betekend een gelijke SE. Voor A is de SE $2/\sqrt{16} = 0.5$. Voor B is \sqrt{n} dus gelijk aan 2; $n = 4$
7. B; ze wil weten of de proporties compenseerders gelijk zijn in de populatie langevluchtmakers en de populatie kortevluchtmakers.
8. A; het zijn onafhankelijke steekproeven en de populatievarianties zijn onbekend dus moet er een t-toets voor onafhankelijke steekproeven gebruikt worden gebruikt worden
9. D; mbt A: een t-toets gebaseerd op gigantisch grote steekproeven zal een kleine p-waarde opleveren ook al zijn de steekproefgemiddeldes nagenoeg gelijk terwijl een t-toets gebaseerd op hele kleine steekproeven best een grote p-waarde kan opleveren terwijl er toch een heel behoorlijk verschil tussen de gemiddeldes zit; mbt B: in het eerste geval zal het BHI vlak bij 0 liggen en in het tweede geval veel verder van 0 af; mbt C een p-waarde zegt niet direct iets over de waarschijnlijkheid van de nulhypothese. Zelfs wanneer je data met behulp van de computer simuleert (en je dus met 100% zekerheid weet dat de nulhypothese waar is), kan je nog een kleine p-waarde vinden.
10. A: de p-waarde bereken je met behulp van de binomiale verdeling (of met behulp van de appendix op blz 646): $H_0: \pi = 0.5$: tabel IIIc levert als eenzijdige p-waarde 0.011 en dus als tweezijdige 0.022. Dit is significant. De hypothese van de onderzoekster was echter dat personen eerder geneigd waren om te compenseren op korte vluchten en dat vinden we ook (in tegenstelling tot het resultaat van opgave 8)
- 11 C; kijk bij de t-test!
- 12 D; ze heeft nu per persoon twee getallen en kan een verschilscore berekenen en zo ieder persoon als zijn eigen controle gebruiken.
- 13 B; als je dit doet zal je kans op een type I fout ongeveer 2 maal zo groot zijn als de α die je kiest!
- 14 B; de formule wordt dus: bedrag = $2 \cdot (0.18/1000) \cdot 13 \cdot \text{reisafstand} + 1.5$. Er is geen errorterm, dus de correlatie is 1 en de residuele variantie (variantie van de errorterm) is $0.2 \cdot (0.18/1000) \cdot 13$ is positief; maar het intercept (wat betaal je bij een reisafstand van 0) is gelijk aan de administratiekosten en dus niet 0.
- 15 A; zie blz 480 van het boek; het interval van C ziet in de buurt maar je kan aan de symmetrische vorm al zien dat het niet klopt (alleen betrouwbaarheidsintervallen bij een gevonden correlatie van 0 zijn symmetrisch!).
- 16 D; Jaarinkomens in euro's zijn vrij groot (~10000 tot 200000), de hypothetische compensaties zijn veel kleiner (~10 tot ~100); om dit passend te krijgen moet het regressiegewicht van inkomen in de orde van 0.001 tot 0.0001 liggen (het geschatte gewicht was 0.0004187057243593 om precies te zijn) en met slechts 3 cijfers achter de komma kom je dan al snel op 0.000. Maar de significantie geeft al aan dat inkomen toch echt in het model hoort!
- 17 B het regressiegewicht van compensatie is significant en negatief. Het gemiddelde hypothetische bedrag is in de WEL groep dus lager dan in de NIET groep en dus worden

kortere reizen vaker gecompenseerd dan langere reizen. De interactie is duidelijk niet significant en speelt dus geen rol.

18 D en dan vooral plaatjes met de verdeling van de residuen (voor de assumptie van normaliteit) en de residuen als functie van de voorspelde waarde (voor de rest).

19 C