**Statistiek KW/MBW**

( tweede deel; tweede gelegenheid )

|  |  |
| --- | --- |
| 1. | Alle antwoorden dienen gemotiveerd te worden. |
| 2. | Alle antwoorden dienen afgerond te worden op *vier* decimalen, tenzij anders vermeld. |
| 3. | Het raadplegen van het theorieboek “Statistiek om mee te werken” (A. Buijs) en het bijbehorende register is toegestaan. |
| 4. | De aanwezigheid van *communicatieapparatuur* is niet toegestaan. |
| 5. | Het gebruik van een (grafische) rekenmachine met statistische programmatuur en het raadplegen van de bijbehorende handleiding is toegestaan. Het *statistische* gebruik van deze rekenmachine is bij een aantal onderdelen beperkt. Let op de aanwijzingen! Het programmeerbare deel mag geen informatie bevatten, die betrekking heeft op de collegestof. |
| 6. | De opgaven dienen na afloop van het tentamen ingeleverd te worden. |

# Opgave 1

Een machine vult pakken met suiker. De volgende gewichten x worden door ons waargenomen: 503, 496, 510,

504, 492, 496, 503. Hierbij nemen wij aan dat de gewichten x normaal verdeeld zijn met een gegeven σ = 6.

1. Bereken – zonder gebruik te maken van de optie TESTS van de grafische rekenmachine - een 90%betrouwbaarheidsinterval voor μ op grond van bovengenoemde steekproef.

Betrouwbaarheidsinterval met betrouwbaarheid voor bij gegeven waarbij geschat wordt met het steekproefgemiddelde van meetwaarden: Bepaal eerst de -waarde bij een rechteroverschrijdingskans van . Het gevraagde interval is dan .

Bij hoort bij een rechteroverschrijdingskans de waarde dus interval

1. Het bij onderdeel a) bepaalde interval mag hoogstens 5 lang zijn. Hoe groot dient de steekproefomvang minstens te zijn om aan deze eis te voldoen?

De breedte van het interval is . Hieruit volgt is . Naar boven afronden voor alle zekerheid: .

1. Toets H0: μ = 505 tegen H1: μ < 505. Bepaal de toetsuitslag via het berekenen van een kritiek gebied. Kies hierbij α = 0,05.

De toetsingsgrootheid is normaal verdeeld met gemiddelde μ = 505 en standaarddeviatie .

Het kritieke gebied is met eld . Nu ligt in dit kritieke gebied, dus wordt verworpen.

1. Bepaal – zonder gebruik te maken van de optie TESTS van de grafische rekenmachine - de toetsuitslag van onderdeel c) via het berekenen van een overschrijdingskans. Kies opnieuw α = 0,05.

Veronderstel weer dat de toetsingsgrootheid is normaal verdeeld met gemiddelde μ = 505 en standaarddeviatie . De p-waarde =

Deze waarde is kleiner dan de significantie α = 0,05, dus wijkt significant af van de aangenomen . wordt dus verworpen.

# Opgave 2

Als gevolg van het referendum over het associatieverdrag met Oekraïne hebben mensen meer belangstelling voor krantenartikelen met betrekking tot het Europese parlement. Ook het leesgedrag van adelborsten en cadetten is hierbij in kaart gebracht. Onderstaande tabel met 80 adelborsten en cadetten afkomstig uit CLAS, CLSK en CZSK bevat informatie met betrekking tot het aantal uren dat per week aan het lezen van dergelijke artikelen wordt besteed. We veronderstellen dat er sprake is van een steekproef.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0 - < 1 | 1 - < 3 | ≥ 3 | Totaal |
| CLAS | 16 | 14 | 10 | 40 |
| CLSK | 6 | 6 | 3 | 15 |
| CZSK | 5 | 10 | 10 | 25 |
| Totaal | 27 | 30 | 23 | 80 |

1. Ga via een homogeniteitstoets na of er sprake is van samenhang tussen *krijgsmachtdeel* en *belangstelling voor het Europese parlement*. Formuleer eerst een nulhypothese en een alternatieve hypothese. Bepaal de uitslag van de toets via een kritiek gebied. Kies α = 0,05.

Een homogeniteitstoets is de toets. Standaard voor deze toets is de formulering:

: Geen samenhang.

: Wel samenhang.

Doe de toets: Bereken eerst de verwachte frequenties uit de som- en rijtotalen op grond van de aanname () dat rijen en kolommen niet van elkaar afhankelijk zijn. Dan geldt bijvoorbeeld dat een gedeelte 40/80 in CLAS zit en gedeelte 27/80 dat maximaal een uur leest, dus in CLAS verwacht je (40/80)⋅(27/80)⋅80 = 13,5 personen die maximaal een uur lezen:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0 - < 1 | 1 - < 3 | ≥ 3 | Totaal |
| CLAS |  |  |  | 40 |
| CLSK |  |  |  | 15 |
| CZSK |  |  |  | 25 |
| Totaal | 27 | 30 | 23 | 80 |

Bereken de toetsgrootheid:

Deze toetsgrootheid voldoet aan een -verdeling met aantal vrijheidsgraden: (3-1)\*(3-1) = 4.

In Tabel 7 vind je de grenswaarde . Het kritieke gebied is dus: .

De berekende waarde van ligt niet in , dus H0 wordt niet verworpen; er is dus geen sprake van samenhang.

Eén van de berekende verwachte frequenties is kleiner dan 5, in dat geval werkt de -toets mogelijk niet goed. Er zijn twee mogelijke aanpakken om dat te verhelpen:

1. rijen 1 en 2 samenvoegen; = 3,6984; ; verwerp H0 niet; óf:
2. kolommen 2 en 3 samenvoegen; = 3,0748; ; verwerp H0 niet.

De conclusie (Er is geen samenhang) geldt dan wel onder die specifieke aanname die hoort bij dat samenvoegen, bv. in het eerste geval moet bij “krijgsmachtdeel” CLAS en CLSK bij elkaar als een deel worden beschouwd en zou het kunnen zijn dat als je ze als aparte delen beschouwt er binnen CLAS en CLSK heel verschillende verdelingen gelden (bv. binnen CLASS leest helemaal niemand), waardoor de algemene conclusie niet meer zou gelden

Veronderstel dat het totaal aantal officieren CLAS in actieve dienst gelijk is aan 3600, het aantal officieren CLSK in actieve dienst gelijk is aan 1200 en dat dit aantal bij CZSK gelijk is aan 2400.

1. Is genoemd onderzoek representatief wat betreft *krijgsmachtdeel*? Beantwoord de vraag via het uitvoeren van een aanpassingstoets. Bepaal de uitslag van de toets via een kritiek gebied. Kies α = 0,05.

Er hebben 80 officieren deelgenomen. Op grond van de gegeven totale aantallen officieren zijn de verwachte aantallen deelnemers aan de toets als deze representatief zou zijn  
CLAS 3600/(3600+1200+2400)\*80 = 40

CLSK 1200/7200\*80 = 13,333

CZSK 2400/7200\*80 = 26,666

Bereken de toetsgrootheid:

De grenswaarde (Tabel C7, 3-1 = 1 vrijheidsgraden) is 5,99

Kritiek gebied

ligt niet in Z, dus wordt niet verworpen, dus de steekproef is wat betreft krijgsdeel representatief.

# Opgave 3

Wanneer de coopertestgegevens van twee opkomstjaren nader bekeken worden, dan ontstaat, ondanks het feit dat de gegevens gecorrigeerd zijn voor geslacht, de indruk dat de scores van de vrouwen achterblijven bij die van de mannen. Om het genoemde verschil nader te onderzoeken beperkt men zich tot de gegevens van cadetten van het dienstvak van de logistiek en de adelborsten van de logistieke dienst. Voor deze groep zijn de gegevens zoals in onderstaande tabel vermeld. We veronderstellen dat er sprake is van een steekproef en dat de variabele *coopertest* het patroon van een normale verdeling volgt.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Steekproefgemiddelde** | **Steekproefstandaarddeviatie** | **Aantal** |
| **Man** | 2590,17 | 117,17 | 9 |
| **Vrouw** | 2325,65 | 144,95 | 11 |

1. Bepaal – zonder gebruik te maken van het menu TESTS van de grafische rekenmachine - een 95%betrouwbaarheidsinterval voor de parameter σ2Man.

Zie Buijs/Reader Par. 10+.1. Het gevraagde betrouwbaarheidsinterval wordt gegeven door

Waarbij je de grenswaarden vindt in Tabel C7 (-verdeling).

Hierbij is ,

,

,

,

.

Het gevraagde interval is dan:

1. Laat via het uitvoeren van een toets zien dat de twee groepen gekenmerkt worden door een even grote variantie. Geef duidelijk aan wat hypothesen, toetsingsgrootheid en kritieke gebied zijn. Kies α = 0,05.

Voor het kijken of twee groepen een gelijke variantie hebben gebruik je de F-toets (Par. 11.5). Hiervoor gelden standaard de volgende hypotheses:

De varianties voor man en vrouw zijn gelijk.

De varianties voor man en vrouw zijn verschillend.

De toetsingsgrootheid is en de waarde is . De vrijheidsgraden van de F-verdeling zijn en .

Het kritieke gebied bestaat uit de vereniging van een linker- en een rechterinterval, beide met een overschrijdingskans van .

De grenswaarde van het rechterinterval zoek je op in Tabel 8 (want overschrijdingskans is 0,025) bij zijn en , je vindt dan en rechterinterval .

Het linkerinterval kun je niet direct opzoeken, maar wel met een truc, want omdat

kun je de **linkergrens** uitrekenen als **1 gedeeld door de rechtergrens**, waarbij je de twee vrijheidsgraden van volgorde verwisselt. Je zoekt dus in Tabel 8 bij en (volgorde verwisseld) en vindt 4,30. De linkergrens is dan 1/4,30 = 0,2326 met interval

Het kritieke gebied is dus .

De berekende waarde ligt niet in een van de twee intervallen, dus niet in het kritieke gebied. Dat betekent dat niet wordt verworpen dus dat waar is, d.w.z., de varianties voor mannen en vrouwen zijn statistisch significant (met α = 0,05) gelijk.

1. Scoren mannen bij de coopertest significant *hoger* dan vrouwen? Voer een toets uit. Geef duidelijk aan wat hypothesen, toetsingsgrootheid en kritieke gebied zijn. Kies α = 0,05.

Je wilt weten of mannen beter scoren, dus je moet kijken naar de gemiddelde scores:

.

.

De varianties zijn niet bekend. Als je van het gegeven van onderdeel b gebruik maakt dan zijn de varianties van de groepen gelijk en is hier de t-toets uit Par. 11.2.2 toepasbaar zijn (als je daar niet vanuit gaat moet je de toets uit Par. 11.2.3 toepassen

De toetsingsgrootte is , die voldoet aan een t-verdeling met vrijheidsgraden.

In de formule van is het verschil van de gemiddelde score van mannen en vrouwen 264,52.

Omdat we aannemen, geldt .

.

De **pooled variance** .

Er volgt dan dat

Bij α = 0,05 enkelzijdig en 9+11-2 = 18 vrijheidsgraden vind je in Tabel C6 een grenswaarde van 1,734, dus het kritieke gebied is . De berekende waarde van t zit hierin, dus wordt verworpen, dus mannen scoren significant beter dan vrouwen.

1. Hoe groot dient het verschil tussen mannen en vrouwen minimaal zijn om significant te zijn?

De grenswaarde voor die nog net toelaatbaar is bij de gekozen significantie is . Er geldt dan

dus

is de kleinste waarde van het verschil .

# Opgave 4

In zes filialen van Bruna boekwinkels is bijgehouden hoeveel exemplaren van het populaire voetbalboek “Gijp” van journalist Michel van Egmond werden verkocht in de week vóór de aanvang van het WK 2014 èn in de week erna. De resultaten staan vermeld in onderstaande tabel:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Filiaal** | **1** | **2** | **3** | **4** | **5** | **6** |
| **Vóór aanvang WK 2014** | 12 | 7 | 28 | 20 | 18 | 11 |
| **Ná aanvang WK 2014** | 22 | 18 | 27 | 29 | 28 | 20 |

Veronderstel dat er sprake is van een normale verdeling.

1. Bereken de correlatiecoëfficiënt van Pearson.

De correlatiecoëfficiënt van Pearson is een getal tussen -1 en +1 dat aangeeft hoe goed twee variabelen aan een lineair verband voldoen. In dit geval is dat X = de verkoop vóór het WK en Y = de verkoop erna.

De correlatiecoëfficiënt is

Want

De correlatiecoëfficiënt is positief, dus er is een positieve correlatie (d.w.z. bij een grotere verkoop vóór het WK hoort een grotere verkoop erna, her lineaire verband tussen verkoop voor en verkoop na is een rechte lijn die stijgend is.

Hoe dichter bij 1 (of -1), hoe beter de correlatie. In dit geval dus een behoorlijk goede correlatie. Dat betekent dat er een behoorlijk goed lineair verband zal zijn tussen X en Y, dus het is verantwoord om lineaire regressie toe te passen.

1. In een zevende Bruna filiaal zijn in de week vóór de aanvang van het WK 2014 15 exemplaren verkocht. Geef een statistisch verantwoorde voorspelling van het aantal verkochte exemplaren in dit filiaal in de week erna.

De regressielijn is met

Vul verkoop vóór WK in: in en je krijgt een bijbehorende voorspelling na WK van .

1. Beantwoord via het uitvoeren van een gepaarde t-toets de vraag of er sprake is van een verkooptoename als gevolg van de start van het WK 2014. Bepaal de toetsuitslag via het berekenen van een kritiek gebied. Kies hierbij α = 0,05.

**Berekening via het kritieke gebied:**

Bij de t-toets wordt bij elk paar gekeken naar het verschil en wordt gekeken of de gemiddelde waarde van deze verschillen 0 is, of afwijkt van 0:

: Geen verband, d.w.z.

: Positief verband, een stijging, d.w.z.

α = 0,05 betekent dat de kans dat je accepteert terwijl eigenlijk waar is maximaal 0,05 mag zijn.

De toetsingsvariabele is

en gedraagt zich volgens een t-verdeling met *n-1* vrijheidsgraden.

Zoek de grenswaarde voor de t-verdeling met α = 0,05 en 6-1=5 vrijheidsgraden op in Tabel C6: , dus kritieke gebied is . De berekende waarde van ligt hierin, dus wordt verworpen.

**Berekening m.b.v. p-waarde:**

Bereken de overschrijdingswaarde: Bij

en 6-1=5 vrijheidsgraden vind je een rechteroverschrijdingswaarde van

Deze waarde is kleiner dan α = 0,05, dus wordt verworpen, wordt aangenomen, dus er is een significante stijging van verkoopwaarden door het WK.