Introductie SPSS

Hogeschool Gent

Mei 2010

Inhoudsopgave

1 Introductie **2**

2 One-sample T-test **11**

3 Two-sample T-test **14**

4 Paired T-test **18**

5 Anova **21**

6 Categorische Data-analyse **26**

6.1 Binomiaaltoets 26

6.2 Pearson χ^2-toets 28

7 Enkelvoudige lineaire regressie **30**

1 Introductie

1. **SPSS openen**
   Klik op het icoontje `SPSS`. Wanneer je het programma opstart krijg je een venster, dit mag je sluiten door op `[cancel]` te drukken. De SPSS data editor staat nu op het scherm. Je kan drie delen onderscheiden:
   1. Het bovenstuk of de menubalk zoals bij Windows de gewoonte is.
   2. Het middenstuk of de data editor, hier voer je je data in.
   3. Het onderstuk of de status balk, hier krijgt de gebruiker informatie omtrent de status van SPSS (welke opdrachten uitgevoerd worden).

2. **Invoer van data**
   De kolommen van je data editor representeren variabelen, de eigenschappen van mensen, objecten,... die je gemeten hebt. De rijen stellen cases voor, proefpersonen, elementen in je steekproef.
   Als een oefening vullen we het gewicht en de lengte van tien personen in.

gewicht	lengte	naam
60	1,68	Pieter
75	1,70	Sofie
90	1,85	Els
95	1,87	Karen
73	1,80	Jan
62	1,70	Lies
57	1,65	Leen
84	1,85	David
80	1,90	Jeroen
78	1,79	Thomas

   We nemen de variabele 'gewicht' en 'lengte' en vullen daar waarden van in. De variabele moet natuurlijk ook een naam krijgen want met de benaming die SPSS geeft, namelijk `[var00001]` ben je niets. Een naam geven aan een variabele kan op twee manieren:

   a) Je gaat naar het tabblad `variable view` en vult daar de naam in.

b) Je dubbelklikt op het grijs vakje (waarin de naam [var00001] staat) in je data editor en komt zo automatisch op het tweede tabblad terecht.

We noemen deze eerste variabele 'gewicht' en de tweede variabele 'lengte'.
De derde variabele die we gemeten hebben zijn de namen van onze cases. Die vullen we in, maar SPSS toont de ingevoerde 'waarden' niet. Dit komt omdat het programma alle ingevoerde waarden als numeriek ziet, als getallen. Wanneer dan letters ingevoerd worden -zoals namen- herkent SPSS dit niet, tenzij je expliciet meedeelt dat je letters gaat invoeren. Dit moet je opnieuw in het tabblad `variable view` gaan specifiëren. Dit tabblad is heel handig omdat je overzicht krijgt van de eigenschappen van je ingevoerde variabelen.
We geven een naam aan onze derde variabele, 'naam'. Het type van deze variabele zal niet numeriek zijn, maar een woord of een letter. In de informaticawereld gebruikt men de benaming 'string' voor een aaneenschakeling van karakters, hetzij letters, hetzij cijfers of een combinatie van letters en cijfers. Wanneer je 'string' aangevinkt hebt dan kan je opnieuw naar de data editor gaan en daar je gegevens invullen.
Stel nu dat we de naam voor onze eerste variabele 'gewicht' willen veranderen naar 'weight'. Dat kan opnieuw via het tabblad `variable view`, daar klik je op de naam van je variabele, je vult een nieuwe naam in en bevestigt.
We hebben nu tien cases en drie variabelen. We willen echter de eerste case of het eerste element eruit laten. Via een klik (met de rechtermuisknop) op het grijze vakje links van de rij die je wil verwijderen, en door vervolgens `clear` te selecteren, verwijder je het element. Dezelfde werkwijze om een variabele te verwijderen. Een volledige rij of variabele selecteren kan ook.
Variabele-namen in SPSS mogen maximaal 8 karakters hebben. Dat wil zeggen dat een naam als 'gewicht in kilogram' niet zal aanvaard worden wegens te lang. In je document kan je de naam van de variabele korter maken en tegelijk een label voor je variabele aanmaken. Via het tabblad `variable view` onder de kolom label kan je verduidelijken wat je met de naam van je variabele bedoelt.

3. **Document opslaan**
   `File - Save as...  voorbeeld`
   `Beeld - Opties`
   De extensie voor SPSS bestanden is `.sav`, net zoals je bij Word de extensie `.doc` hebt.

4. **Bestanden importeren**

   In SPSS kan je bestanden openen van verschillende types. Vanuit Excel kan je gegevens gaan importeren naar SPSS. Een voorbeeld:

   a) Start Excel op

   b) Voer volgende tabel in

1	5
2	4
3	3
4	2
5	1

   c) Sla het bestand op

   d) `SPSS: File - New Data`

   e) `SPSS: File - Open Data`

   f) Zoek het .xls bestand. Afhankelijk van de complexiteit van het .xls bestand zal SPSS meerdere zaken vragen vooraleer het bestand te importeren. In dit geval hebben we te maken met een basisbestand, een druk op de `Enter`-toets volstaat.

   g) Nog gemakkelijker gaat het via `Copy` en `Paste`

   Ook in Wordpad of Kladblok kan je je data in een .txt bestand opslaan en dit gaan importeren in SPSS.

5. **Het tabblad variable view**

   a) NAME

   Naam van de variabele. Mag maximaal acht karakters hebben.

   b) TYPE

   SPSS ziet de ingevoerde gegevens standaard als numeriek.

   Je hebt nog andere types zoals:

→ numeric	12,5
→ comma	12,5
→ dot	12.5
→ scientific notation	$12.5E03 = 12500$
→ data	12/05/2002
→ dollar	
→ custom currency	
→ string	twaalf

c) WIDTH
De breedte van de waarde voor de variabele (aantal cijfers of letters).

d) DECIMALS
Aantal getallen na de komma bij een numerieke variabele.

e) LABEL
Beschrijving van de variabele.

f) VALUES
bij categorische variabelen kan je codes toekennen aan de categorieën. Neem de variabele `[geslacht]` met de waarden 'man' en 'vrouw'. We kunnen nu telkens die strings gaan invullen of, en wat beter is, we kunnen een code toekennen aan elk van die twee waarden, bvb. aan 'man' de code 2, aan 'vrouw' de code 1. (In de datamatrix vullen we dus de scores '1' en '2' in.) Vervolgens kunnen we in het tabblad variable view (onder de kolom `values`) definiëren waarvoor beide scores staan.

```
Klik [...]
value:  2 - value label:  man - [Add]
value:  1 - value label:  vrouw - [Add]
[OK]
```

Wat gebeurt er wanneer je volgende actie vraagt?

```
VIEW - Value Labels
```

g) MISSING
Stel dat je van een bepaalde persoon een waarde tekort hebt. Van case nummer vijf weten we de leeftijd niet. Dit noemen we een 'missing value'.

Een concreet voorbeeld: De meeste studenten hebben het examen Statistiek gemaakt en ingediend. We voeren hun punten in. Van een paar studenten hebben we echter geen punten, dit komt omdat ze ofwel pro forma hebben ingediend of omdat ze ziek waren. Er is een verschil tussen beide redenen en dit kunnen we specifiëren in het tabblad `variable view`. Je kan werken met discrete missing values, vb. de code 98 voor een student die pro forma indiende, de code 99 voor een zieke student. Probleem bij SPSS is dat je die codes voor missing values geen label kan geven. Je kan ook een range specifiëren.

h) COLUMNS

i) ALIGN
Standaard worden numerieke variabelen rechts uitgelijnd en worden string-variabelen links uitgelijnd.

j) MEASURE
Er zijn vijf meetniveaus:

nominal	vb. geslacht
ordinal	vb. plaats in een wedstrijd
scale: interval-ratio-absoluut	vb. temperatuur

6. **Demonstratie SPSS**

Data: `voorbeeld.sav`

a) `Analyse - Descriptive Statistics - Frequencies`
Hiermee krijg je de beschrijvende statistieken van de variabelen die opgegeven worden. In het linkerkader staan alle variabelen uit de dataset, het rechterkader bevat de variabelen waarvan de frequenties worden gegeven. Door in het linkerkader een variabele aan te klikken, en vervolgens op de zwarte pijl te drukken (waardoor de variabele naar het rechterkader verplaatst wordt), geef je aan de je van respectievelijke variabele de beschrijvende statistieken wil krijgen.
→ Vraag de frequenties op van de variabele `[leeftijd]`
Er verschijnt een output venster. Let op het verschil tussen ‘valide percent‘ en ‘percent‘. Onder de kolom ‘valid percent‘ staan de waarden als we rekening

willen houden met de missing values in de dataset. Bij het berekenen van dit 'valide percent' worden de missing values virtueel verwijderd.
In de output staan de kaders of tabellen als apart object. In het linkerdeel van het outputvenster zie je een overzicht van de inhoud van de output. De kaders kan je gemakkelijk in een ander document (vb. Word) gaan invoegen door te klikken op de rechtermuisknop en `Copy as object` te selecteren. Als je dubbelklikt op een kader dan verschijnt de Toolbox, hiermee kan de layout van de kaders aangepast worden.

STATISTICS

Onder die knop kan je diverse maten aanvinken:

maten van centrale tendentie	Central Tendency
maten van spreiding	Dispersion
percentielen	Percentile Values
verdeling	Distribution

[Opmerking: kurtosis doelt op de mate van welving, scheefheid of 'skewness' doelt op de mate waarin de verdeling scheef is. Verder informatie kan je vinden in de Help-functie van SPSS of in handboeken Statistiek.]

CHARTS

`Chart Type: histogram`

`v With normal curve`

Je kan een histogram opvragen van variabelen en direct de normaalverdeling erbij vragen. Zo valt er op het eerste gezicht te zien of een verdeling al dan niet normaal verdeeld is. Statistisch meer onderbouwd is er echter de Kolmogorov-Smirnov toets waarmee je kan toetsen of een verdeling al dan niet normaal is.

`Analyse - Descriptive Statistics - Explore`

`plots...`

`v Normality plots with tests`

SPSS voert de standaard analyses uit `[case processing summary], [descriptives], [test of normality]` en geeft een heleboel informatie waaronder betrouwbaarheidsintervallen, boxplots enzoverder.

GRAPHS

Kan je aanvinken bij meerdere vensters. Een andere manier is verder in dit document beschreven.

b) `Analyse - Descriptive Statistics - Descriptives`
Dit onderdeel bevat ongeveer dezelfde informatie als
`Analyse - Descriptive Statistics - Frequencies`,
behalve dat onder de knop OPTIONS de volgorde van verschijnen kan aangepast worden.

c) `Graphs`
`Graphs - Legacy - Boxplot`
`Simple - v summaries of separate variables - Define -`
`[inkmoe]`
`Simple - v groups of cases - Define -`
`variable: [inkmoe] - category axis [geslacht] - label cases by`
`[naam]`

[Opmerking: het volle vlak in de figuur representeert de interkwartielafstand, de bovenste en onderste horizontale lijn geven de afstand aan waarbinnen cases moeten vallen als ze niet als outlier beschouwd willen worden. Alles boven de bovenste horizontale lijn en onder de onderste horizontale lijn kan men beschouwen als outlier.]

• INTERVALNIVEAU
Hebben we te maken met variabelen van intervalniveau of hoger dan maken we een scatterplot. Die kunnen we op twee manieren opvragen: (1) via de plots knop in een van de vensters of (2) via het menu `Graphs`.
Als we de relatie willen nagaan tussen het inkomen van de moeder en de afstand tot school kunnen we een scatterplot vragen aangezien we te maken hebben met twee variabelen van minstens intervalniveau.
`Graphs - Legacy - Scatter`
`Simple - Define -`
`Y axis: afstand tot school - X-axis: inkomen moeder`
We krijgen opnieuw een outputvenster waarin de scatterplot afgebeeld staat. De maat voor samenhang is hier de Pearson Correlatiecoëfficiënt (default). De correlatiecoëfficiënt gaat van -1 tot +1. Helemaal negatief betekent dat er

een sterk negatief of dalend verband is. Een getal nul wil niet zeggen dat er geen verband zou zijn. Er kan een curvilineair verband zijn. Daarom is het ook altijd nodig om een scatterplot te maken want daar is op te zien hoe je de waarde van de correlatiecoëfficiënt kan interpreteren. De correlatiecoëfficiënt zelf kan je opvragen via

`Analyse - Correlate - Bivariate`

Het outputvenster geeft je de waarden voor de correlatiecoëfficiënt.
Wanneer je van meer dan twee variabelen de correlatie wil opvragen krijg je een grotere tabel in de output.

[Opmerking: Het *-teken naast de waarden in een tabel betekent dat die waarden significant zijn op het 5%-niveau. Als een verband significant is op het 1% niveau dan staat er ** naast de waarde.]

d) `Extra`

• Variabelen hercoderen

Een variabele hercoderen naar een andere [different = nieuw aangemaakte variabele, zodat de oorspronkelijke niet overschreven zal worden] variabele. We kunnen de waarden van de variabele `geslacht` andere meetwaarden geven als volgt: `Transform - Recode - Into different variables...`
`input variable: geslacht - output variable: sexe -` `change`
`Old and New Values`
`Old value: 1 - New value: 2 -` `Add`
`Old value: 0 - New value: 1 -` `Add`

[Opmerking:Dit ziet er op het eerste zicht totaal zinloos uit maar is wel populair in gebruik. Stel je de variabele [loon] (van intervalniveau) voor van zo'n 1000 cases. Een manier om deze variabele overzichtelijk te maken is in categorieën in te delen. Dit kan via het voorgaande commando waar een bepaalde range een code toegekend krijgt.]

• Variabelen berekenen

We willen het totaal aantal vervoermiddelen berekenen waarmee iemand naar school komt.

`Transform - Compute`
`target variable: vervoer`

`Numeric expression: auto+trein+voet+fiets+tram+bus`
Er staan ook een aantal standaardfuncties in het venster.

• Cases selecteren
We willen alleen die cases betrekken waarvan de moeder een inkomen heeft dat hoger is dan 50000. `Data - Select Cases`
`if... condition is satisfied -` `if...`
`‘inkmo > 50000‘ invullen`
In de dataset merk je dat bepaalde cases weggelaten zijn door de schuine streep die getrokken is.

2 One-sample T-test

Een onderzoeker wil nagaan of de gemiddelde inname van calorieën bij vrouwen verschilt van de vooropgestelde waarde 7725. Van tien vrouwen wordt de calorie-inname gemeten, hun waarden staan hieronder.

subject	inname
1	7000
2	6800
3	8500
4	5600
5	5800
6	6350
7	8230
8	8100
9	5400
10	5980

Opgave

Ga na of de gemiddelde calorie-inname van deze groep vrouwen significant verschilt van de gemiddelde calorie-inname in de populatie, namelijk 7725. Gebruik $\alpha = 0.05$. Bereken ook het gepaste betrouwbaarheidsinterval.

Oplossing

Om na te gaan of er een verschil is tussen het geobserveerd gemiddelde en het populatie-gemiddelde zullen we gebruik maken van een t-toets voor één steekproef. De nulhypothese kan men als volgt formuleren: $H_0 : \mu = 7725$, of nog, $H_0 : \mu - 7725 = 0$. De alternatieve hypothese stelt $H_a : \mu \neq 7725$, of nog, $H_a : \mu - 7725 \neq 0$. Het betreft hier dus een tweezijdige toets.

In SPSS gaan we als volgt te werk:

1. **Data invoeren**
   In SPSS representeren de rijen de subjecten, en de kolommen de verschillende variabelen die we van deze subjecten gemeten hebben. Hier is er slechts één variabele, en zullen we dus één kolom met 10 scores moeten invoeren.
2. **Procedure**

   ```
   Analyze --> Compare Means --> One-Sample T Test ...
       Test Variable(s): inname
       Test Value: 7725
       [OK]
   ```

   Bij 'Test Value' vullen we de vooropgestelde waarde in (i.e. 7725).
3. **Output** De tabel `One-Sample Statistics` bevat enkele beschrijvende statistieken: aantal observaties (`N`), geobserveerd gemiddelde (`Mean`), standaardafwijking (`Std. Deviation`) en de standaardfout (van de steekproevenverdeling van het gemiddelde) die we nodig hebben om de t-statistiek te berekenen (`Std. Error Mean`). In de tabel `One-Sample Test` staan de toetsresultaten. De t-statistiek wordt eerst gegeven. Daarna volgt het aantal vrijheidsgraden (dit is hier gelijk aan $n - 1$). Vervolgens wordt de (tweezijdige) overschrijdingskans (i.e. de p-waarde) gerapporteerd in de kolom `Sig. (2-tailed)`. In de kolom `Mean Difference` staat het verschil tussen het geobserveerde gemiddelde en de vooropgestelde waarde. Tenslotte, het betrouwbaarheidsinterval in de laatste twee kolommen is gebaseerd op dit *verschil* (tussen geobserveerd en voorgesteld gemiddelde).Dit is niet gebruikelijk maar men kan eenvoudig het juiste interval vinden door zowel de benedengrens als de bovengrens op te tellen met 7725.

Oefening

Een studiebegeleider veronderstelt dat studenten tijdens het hele jaar gemiddeld drie uur per dag moeten studeren om niet voor verrassingen te komen staan tijdens de examens.

Aantal uren	3.7	2.7	3.3	1.8	4.6	4.0	2.3	2.4	2.1	2.2

Tabel: steekproef van 10 studenten met het gemiddeld aantal studieuren per dag per student.

Doet onze steekproef vermoeden dat studenten gemiddeld minder studeren dan de studiebegeleider vooropstelt? Ga dit na met een betrouwbaarheid van 95%.

3 Two-sample T-test

Opgave

Een onderzoeker wil nagaan of kinderen beter in staat zijn 'concepten' te leren wanneer enkel positieve voorbeelden gebruikt worden in tegenstelling tot een situatie waarin zowel positieve als negatieve voorbeelden gebruikt worden. De kinderen worden at random toegekend aan één van de twee experimentele condities. Hun scores op de leertaak worden hieronder weergegeven.

positief	positief en negatief
8	14
10	8
7	7
12	10
6	12
9	6
10	15
11	11
6	9
13	8

De onderzoeker verwacht dat de kinderen uit de eerste groep (positief) betere scores behalen dan de kinderen uit de tweede groep (positief en negatief). Gebruik $\alpha = 0.01$. Bereken ook het gepaste betrouwbaarheidsinterval.

Oplossing

De kinderen werden at random in twee verschillende groepen ingedeeld. Het gaat hier dus over onafhankelijke steekproeven. Om na te gaan of er een verschil is tussen de gemiddelde scores van de twee groepen zullen we gebruik maken van een t-toets voor twee onafhankelijke steekproeven. De alternatieve hypothese kan men als volgt

formuleren: $H_a : \mu_1 > \mu_2$, of nog, $H_a : \mu_1 - \mu_2 > 0$. Het betreft hier dus een eenzijdige (rechtszijdige) toets.

De geobserveerde gemiddelden bedragen respectievelijk $\bar{x}_{\text{pos}} = 9.2$ en $\bar{x}_{\text{posneg}} = 10$, en dus $\bar{x}_{\text{pos}} - \bar{x}_{\text{posneg}} = -0.8$. In tegenstelling tot de verwachtingen van de onderzoeker is de (gemiddelde) score voor de eerste groep *slechter* dan de tweede groep. Onze alternatieve hypothese kunnen we dus onmiddellijk verwerpen. Een formele toets is in feite overbodig. Niettemin zullen we hieronder formeel de t-toets uitvoeren (om uiteindelijk te concluderen dat we inderdaad de nulhypothese dienen te behouden).

In `SPSS` gaan we als volgt te werk:

1. **Data invoer**
   Binnen `SPSS` krijgen we in elke rij de scores van een welbepaald subject op één of meerdere variabele(n). De scores op eenzelfde variabele komen in eenzelfde kolom. Bovenstaand voorbeeld betreft 20 verschillende subjecten, waarvan we de scores dus onder elkaar moeten invoeren. De aanmaak van een tweede variabele is nodig om de subjecten van de verschillende groepen van elkaar de onderscheiden. De eerste variabele `score` is dus de score op de leertaak, de tweede variabele `groep` is de conditie waartoe de subjecten behoren. Voor deze laatste variabele gebruiken we numerieke codes voor elk van de twee niveaus van de variabele (bvb. '1' voor positief, en '2' voor positief+negatief).

	scores	groep
1	8,00	1,00
2	10,00	1,00
3	7,00	1,00
4	12,00	1,00
5	6,00	1,00
6	9,00	1,00
7	10,00	1,00
8	11,00	1,00
9	6,00	1,00
10	13,00	1,00
11	14,00	2,00
12	8,00	2,00
13	7,00	2,00
14	10,00	2,00
15	12,00	2,00
16	6,00	2,00
17	15,00	2,00
18	11,00	2,00
19	9,00	2,00
20	8,00	2,00
21		
22		

2. **Procedure**

Om de t-toets voor twee onafhankelijke steekproeven uit te voeren in SPSS gaan we als volgt te werk:

```
Analyze --> Compare Means --> Independent Samples T-Test...
   Test Variable(s): score
   Grouping variable: groep
   [Define Groups...]
       Group 1: 1
       Group 2: 2
       [Continue]
   [Options...]
       Confidence Interval: 99
       [Continue]
   [OK]
```

Bemerk:
(1) dat we het betrouwbaarheidsinterval hebben aangepast naar 99 procent,
(2) dat we nergens de mogelijkheid krijgen om een eenzijdige toets te bevragen. SPSS zal altijd een tweezijdige p-waarde (en bijhorend betrouwbaarheidsinterval)

vermelden. Dit zal uiteraard gevolgen hebben voor de interpretatie van de p-waarde die `SPSS` rapporteert.

3. **Output**

   In de tabel `Independent Samples Test` staan de toetsresultaten. Eerst krijg je de waarde van de Levene-toets. Deze toets gaat na of de varianties in beide populaties homogeen zijn ($s_1^2 = s_2^2$). Deze nulhypothese mag hier aanvaard worden ($p = 0.535$ en dus groter dan α). We kijken vervolgens naar de waarden op de bovenste lijn (*Equal variances assumed*; indien de Levene-toets significant is kijken we verder op de 2de lijn). De t-statistiek bedraagt $t(18) = -.657$ met als (tweezijdige) p-waarde $p = 0.520$. Het (tweezijdig) betrouwbaarheidsinterval bedraagt [-4.307; 2.707].

   De vraagstelling was echter eenzijdig geformuleerd. We moeten een eenzijdige p-waarde berekenen op basis van de gerapporteerde tweezijdige p-waarde. In ons voorbeeld ligt het geobserveerd verschil *niet* in de verwachte richting. De eenzijdig p-waarde is dus gelijk aan $1.0 - (0.520/2) = 0.74$ (indien het verschil wel in de verwachte richting ligt, dienen we de p-waarde enkel te delen door twee). Deze p-waarde ligt ver boven het vooropgesteld significantieniveau ($\alpha = 0.01$). We moeten de nulhypothese behouden.

Conclusie en rapportering

Een t-toets voor twee onafhankelijke steekproeven werd uitgevoerd om na te gaan of de scores van de eerste groep (enkel positieve voorbeelden) hoger lagen dan de scores van de tweede groep. In tegenstelling tot de verwachtingen lag het geobserveerd gemiddelde voor de eerste groep (M=9.2, SD=2.44) lager dan dat voor de tweede groep (M=10, SD=2.98). De t-toets was uiteraard niet significant, $t(18) = -0.657$, $p = 0.74$.

4 Paired T-test

Een psycholoog is geïnteresseerd in de relatie tussen stress en korte termijn geheugen. Hiertoe test hij 10 subjecten voorafgaand aan een situatie waarin stress geïnduceerd wordt. Na de stress-situatie worden de subjecten hertest. De resultaten op de geheugentaak zien er als volgt uit:

prestress	poststress
12	11
14	14
10	8
14	15
14	11
17	14
16	16
11	9
12	11
16	13

Opgave

De onderzoeker verwacht dat de scores bij de poststress meting slechter zullen zijn dan deze van de prestress meting. Gebruik $\alpha = 0.05$. Bereken ook het gepaste betrouwbaarheidsinterval.

Oplossing

Het betreft hier twee metingen (pre en post) van dezelfde personen. Het gaat hier dus om herhaalde metingen en dus afhankelijke steekproeven. Om na te gaan of er een verschil is tussen de gemiddelde scores van de pre- en posttest maken we gebruik van een t-toets voor twee afhankelijke steekproeven. De alternatieve hypothese kan men als volgt formuleren: $H_a : \delta > 0$, immers als de tweede meting slechtere resultaten moet

opleveren, dan zal het 'verschil' tussen de eerste en tweede meting gemiddeld positief zijn. Het betreft hier dus een eenzijdige (rechtszijdige) toets.

In SPSS gaan we als volgt te werk:

1. **Data invoeren**
   Er zijn slechts 10 subjecten (10 cases): we voeren de data van de twee kolommen van de bovenstaande tabel naast elkaar in. De eerste kolom '`pre`' bevat de prestress-metingen, de tweede kolom '`post`' bevat de scores op poststress.
2. **Procedure**
   Uitvoeren van een t-toets voor twee afhankelijke steekproeven:

   ```
   Analyze --> Compare Means --> Paired Samples T Test ...
       Paired Variables: pre post
       [OK]
   ```

   Bemerk dat we in SPSS opnieuw niet de mogelijkheid krijgen om eenzijdig te toetsen.
3. **Output**
   In de tabel `Paired Samples Test` staan de toetsresultaten.
   We kijken naar de t-waarde, $t(9) = 3.096$ met $p = 0.013$ (tweezijdig!). Gezien de richting van het effect in de lijn van de verwachtingen ligt (de post-scores liggen inderdaad gemiddeld lager dan de pre-scores) kan men de eenzijdige p-waarde eenvoudig berekenen door de tweezijdige p-waarde te delen door twee: $p_{\text{eenzijdig}} = 0.0065$. Deze p-waarde ligt onder het significantieniveau. Er is een verschil tussen de twee metingen: we moeten de nulhypothese verwerpen.

Conclusie en rapportering

Een t-toets voor twee afhankelijke steekproeven werd uitgevoerd om na te gaan of de scores van de posttest (M=12.2, SD=2.62) slechter waren dan de scores van de pretest (M=13.6, SD=2.32). De t-toets was significant, $t(9) = 3.10$, $p = 0.006$ (eenzijdig). Deze resultaten zijn conform onze verwachtingen.

Oefening

1. Ga met de One-sample T-test na of er een verschil is tussen `prestress` en `poststress`. Maak hiervoor eerst een nieuwe variabele aan `diff`, waarvoor geldt dat `diff` = `prestress` - `poststress`. Vergelijk de resultaten met de resultaten zoals bekomen met de Paired T-test.
2. Gegeven de resultaten van de geheugentaak. Stel nu dat er twee onafhankelijke groepen getest zijn. Scoren beide groepen even goed op de test? (met $\alpha = 0.05$)

5 Anova

Een studentenclub in Gent gaat een weddenschap aan met een bevriende club uit Leuven. De Gentse studenten leven gedurende twee weken op een exclusief dieet van pizza's, terwijl de Leuvense studenten zichzelf een streng bierdieet hebben opgelegd. Beide clubs vragen zich af door welk dieet men het meest aankomt. Een Brusselse studentenclub hoort van de weddenschap en besluit eveneens de uitdaging aan te gaan. De Brusselse studenten nemen deel aan een streng dieet van friet. Vooraf werd iedereen gewogen en na twee weken diëten werden de verschillen in gewicht geregistreerd.

Onderstaande tabel geeft de steekproefgrootte van de drie condities I: de studenten uit Leuven (1) met het bierdieet, Gent (2) met het pizzadieet en Brussel (3) met het frietdieet, met telkens het verschil in gewicht in kilogram voor en na het dieet (gewicht na - gewicht voor).

subject	bierdieet	pizzadieet	frietdieet
1	1	2	4
2	2	2	3
3	2	3	5
4	2	4	4
5	3	4	4
6	2	4	5
7	1	3	4
8	2	2	5
9	3	3	6
10	3	3	5
	$\bar{y}_1 = 2.1$	$\bar{y}_2 = 3$	$\bar{y}_3 = 4.5$

Het totale groepsgemiddelde bedraagt $\bar{y} = 3.2$

Opgave

Onderzoek of er een verschil is tussen de drie diëten. Gebruik $\alpha = 0.05$:

Oplossing

Om na te gaan of er verschillen zijn tussen de drie dieet-condities (wat betreft het aantal bijgekomen kilo's) voeren we (enkelvoudige) variantie-analyse uit.
Eerst en vooral is belangrijk in te zien dat we hier over twee variabelen beschikken. De eerste variabele –die we voortaan `kilo` zullen noemen– bevat de scores (= het aantal bijgekomen kilo's) van de 30 studenten. De tweede variabele –die we voortaan `dieet` zullen noemen– geeft aan tot welke conditie (i.e. studentenclub) iemand behoort.

- *Oneway ANOVA*
  1. **Data invoer en procedure**
     In `SPSS` dient men de twee variabelen in twee aparte kolommen in te geven. De eerste kolom (`kilo`) bevat de bijgekomen kilo's voor de dertig studenten; de tweede kolom (`dieet`) bevat de groepsindeling (met als codes '1', '2' en '3' voor de drie niveaus van de factor).

     ```
     Analyze -> Compare Means -> One-Way ANOVA ...
         Dependent List: kilo
         Factor: dieet
         [Options] v Descriptive
     [Ok]
     ```

  2. **Output**
     De output van deze analyse vindt men in de `ANOVA`-tabel. Bemerk dat hier het 'klassieke' jargon gehanteerd wordt voor de benoeming van de kwadratensommen. De verklaarde/regressie kwadraten som heet hier de '`Between Groups Sum of Squares`', terwijl de error of residuele kwadratensom hier aangegeven wordt door '`Within Groups Sum of Squares`'. De som van beide geeft ons de '`Total Sum of Squares`'.

- *General Linear Model*
  In wat volgt zullen we in `SPSS` echter een andere procedure volgen, die nauwer aansluit met de 'moderne' visie op variantie-analyse waar men variantie-analyse beschouwt als een speciaal geval van het algemeen lineair model (General Linear

Model). SPSS bevat een '`GLM`' procedure die ons toelaat alle mogelijke lineaire modellen (dus ook klassieke regressie-analyse en klassieke variantie-analyse) op een uniforme wijze uit te voeren. In deze visie beschouwt men `gewicht` als een afhankelijke variabelen, en de factor `dieet` als een predictor van nominaal niveau.

1. **Procedure**
   Voor onze enkelvoudige variantie-analyse gaan we als volgt te werk:

```
Analyze -> General Linear Model -> Univariate ...
    Dependent Variable: kilo
    Fixed Factor(s): dieet
    [Options...]
        Display
          v Descriptive Statistics
          v Parameter Estimates
        [Continue]
[Ok]
```

2. **Output** De resultaten van deze analyse vindt men in de tabellen '`Tests of Between-Subjects Effects`' en '`Parameter Estimates`'. In de eerstgenoemde tabel vinden we op de eerste lijn, naast `'Corrected Model'` de verklaarde kwadratensom ($E_0 - E_p$) en de toets die nagaat of het volledige model (in ons geval met 1 predictor) significant beter fit dan een nulmodel. Daar ons model slechts 1 predictor bevat, is de verklaarde kwadratensom ten gevolge van deze factor (lijn 3, `'DIEET'`) gelijk aan de verklaarde kwadratensom van ons 'volledig' model. Op de lijn van `'Intercept'` vinden we de toets die nagaat of de intercept gelijk is aan nul ($H_0 : \beta_0 = \mu = 0$). In de praktijk is deze toets meestal irrelevant. Op de lijn van `'Error'` krijgen we de error-kwadratensom (E_p). Tenslotte, op de lijn `'Corrected Total'` (en niet `'Total'`!) staat de totale kwadratensom (E_0) vermeldt. (Met `'Total'` bedoelt men in `SPSS` de som van de gekwadrateerde respons-variabelen. Deze waarde is voor ons van geen enkel belang.)

   In de '`Parameter Estimates`' tabel krijgen we de geschatte waarden voor de model-parameters μ, α_1, α_2 en α_3. SPSS maakt bij deze analyse standaard gebruik van een effectenmodel met behulp van GLM-restricties waarbij het laatste niveau van de factor als referentieniveau wordt gehanteerd ($\alpha_3 = 0$). In

de tabel worden tevens t-toetsen gegeven die aangeven of de modelparameters significant zijn. Deze toetsen zijn doorgaans irrelevant, omdat ze betrekking hebben op de afzonderlijke hulpveranderlijken.

De waarden van de modelparameters geven aan wat het verschil is met betrekking tot het referentieniveau. Bijvoorbeeld: het verschil in verwachte gewichtstoename tussen de mensen die een pizza-dieet volgden en diegenen die een friet-dieet doormaakten, is gelijk aan 1.5. (Wat we natuurlijk al wisten op basis van de geobserveerde gemiddelden. Die bevinden zich in de tabel `'Descriptive Statistics'`).

Conclusie en rapportering

In de output vinden we telkens de fameuze `Anova`- tabel, die een overzicht geeft van de kwadratensommen, de '`Mean Squares`' (dit zijn gewoon de kwadratensommen gedeeld door het aantal vrijheidsgraden), en de F-statistiek, die we simpelweg bekomen door de Mean Square corresponderend met het model (of nog, de factor) te delen door de Means Square die correspondeert met de error: $14.7/0.644 = 22.810$. De F-statistiek is significant, want de bekomen overschrijdingskans ($p = 0.000$) is overduidelijk kleiner dan het vooropgestelde signficantieniveau $\alpha = 0.05$. We dienen de nulhypothese te verwerpen: er is wel degelijk een verband tussen de variabelen `dieet` en `kilo`. Of nog: er is een verschil tussen drie dieet-condities wat betreft de bijgekomen kilo's. We kunnen de resultaten als volgt formuleren:

> *Een variantie-analyse werd uitgevoerd met als afhankelijke variabele de bijgekomen kilo's, en met als onafhankelijke variabele de factor* `dieet` *die aangeeft tot welke van de drie dieet-condities de student behoort. Het effect van de factor* `dieet` *was significant,* $F(2, 27) = 22.810$, p*=.000. Er is wel degelijk een verschil tussen de drie dieetcondities.*

Tabel 1
De geobserveerde gemiddelden voor de drie dieet-condities

	bier	*pizza*	*friet*
gewicht	*2.1*	*3*	*4.5*

Oefening

Een farmaceutisch bedrijf wil een nieuw medicijn testen dat duizelingen bij depressieve patiënten zou verminderen. Uit Vlaanderen worden 32 personen met een depressie geselecteerd. Ze krijgen reeds medicatie ter behandeling van hun depresssie. Bijkomend krijgt een deel van de personen het nieuwe medicijn. Er zijn vier groepen. De eerste groep krijgt een placebo in de vorm van suikerwater. De tweede groep krijgt 10 milligram van het nieuwe medicijn toegediend. De derde en de vierde groep krijgen respectievelijk 20 en 30 milligram toegediend. Het bedrijf wil weten wat de optimale dosis is van het nieuwe medicijn om de duizelingen voldoende te laten dalen.
Onderstaande tabel bevat de scores en gemiddelden van de vier condities k: de placebo-groep (1), de conditie met 10 mg (2), de conditie met 20 mg (3) en de conditie met 30 mg (4). Telkens werd per persoon het aantal duizelingen geregistreerd gedurende één week.

placebo	10 mg	20 mg	30 mg
25	23	19	16
27	20	16	17
28	29	21	21
23	27	20	24
21	26	18	23
25	28	19	22
29	29	21	18
28	25	22	25

Onderzoek of er een verschil is tussen de vier condities. Gebruik $\alpha = 0.05$

6 Categorische Data-analyse

6.1 Binomiaaltoets

Een senator weet niet of hij nu voor of tegen een nieuwe milieuwet zou stemmen. Verschillende reeds uitgevoerde opiniepeilingen over het onderwerp geven tegenstrijdige resultaten. De senator beslist om een eigen onderzoek uit te voeren. Als meer dan 60% van alle deelnemers aan het onderzoek de nieuwe wet steunt, zal de senator voor de nieuwe wet stemmen. Uit een random steekproef van 750 kiezers blijkt dat 495 mensen de nieuwe wet steunen.

Opgave

Zal de senator stemmen voor of tegen de wet stemmen? ($\alpha = 0.10$)

Oplossing

Deze vraagstelling heeft betrekking op het vergelijken van twee proporties: een geobserveerde proportie ($495/750 = .66$) en een vooropgestelde proportie $60\% = .60 = \pi_0$. We maken gebruik van de binomiaaltoets. Ofwel hanteren we de 'exacte' methode (gebruik makende van de binomiaalverdeling), ofwel maken we gebruik van de benadering op basis van de normaalverdeling. Gezien de grootte van de steekproef kunnen we perfect de benaderingsmethode hanteren. De nulhypothese stelt $H_0 : \pi = .60$. De (eenzijdige) alternatieve hypothese stelt: $H_a : \pi > .60$.

In SPSS gaan we als volgt te werk:

1. **Data invoeren**
   In ons voorbeeld hebben we twee kolommen in SPSS ingevoerd: de frequentie ('freq') en de code (1=voor, 2=tegen) voor de stemintentie ('stem'). De frequentievariabele zullen we als een 'weging' hanteren via '`Weight Cases`'.

```
Data -> Weight Cases...
     v Weight Cases by:
     Frequency Variable: freq
     [Ok]
```

2. **Procedure**
   In SPSS gaan we als volgt te werk om een binomiaaltoets uit te voeren:

```
Analyze -> Nonparametric Tests -> Legacy -> Binomial...
     Test Variable List: stem
     Test Proportion: .60
     [OK]
```

   De z-score wordt niet vermeld. Enkel de (eenzijdige) overschrijdingskans ('Asymp. Sig. (1-tailed)').

Conclusie en rapportering

Een exacte binomiaaltoets werd uitgevoerd om na te gaan of de geobserveerde proprotie $7/10 = .70$ *(van mensen die voor de nieuwe wet stemmen) groter is dan* 60%*. De gegevens laten niet toe deze alternatieve hypothese te bevestigen,* $P(X \geq 7) = .382$*. De senator zal niet voor de nieuwe wet stemmen (maar had beter een grotere steekproef gebruikt!)*

6.2 Pearson χ^2-toets

Een groot notenverwerkend bedrijf verkoopt pakjes met daarin een notenmengeling. Volgens de verpakking bestaat de mengeling voor 30% uit hazelnoten, 20% walnoten, 20% braziliaanse noten en 30% pindanoten. Een consumentenmagazine vraagt zich af of de bewering op de verpakking wel juist is. Een onderzoeker neemt een aselecte steekproef van 200 noten en telt het aantal noten van elke soort. De resultaten worden hieronder weergegeven:

Soort	Frequentie
Hazelnoten	45
Walnoten	35
Braziliaanse noten	41
Pindanoten	79
Totaal	200

Opgave

Is de bewering op de verpakking juist? ($\alpha = 0.05$)

Oplossing

Hier stelt zich de vraag of de geobserveerde frequenties (of proporties) corresponderen met de vooropgestelde frequenties (of proporties). Hiervoor maken we gebruik van Pearson χ^2-toets.

In SPSS gaan we als volgt te werk:

1. **Data invoeren**
   In ons voorbeeld hebben we twee kolommen in SPSS ingevoerd: de frequentie ([freq]) en een code (1,2,3 of 4) voor de vier soorten ([soort]). De frequentievariabele zullen we als een 'weging' hanteren via '`Weight Cases`'.

```
Data -> Weight Cases...
     v Weight Cases by:
     Frequency Variable: freq
     [OK]
```

2. **Procedure**

```
Analyze -> Nonparametric Tests -> Legacy -> Chi-Square...
     Variable List: soort
     v Values:
        30 [Add]
        20 [Add]
        20 [Add]
        30 [Add]
     [OK]
```

3. **Output**
   In de tabel 'Test Statistics' krijgen we de waarde voor de χ^2-statistiek, het aantal vrijheidsgraden (df) en de overschrijdingskans (p).

Conclusie en rapportering

Een Pearson χ^2-toets werd uitgevoerd om na te gaan of de geobserveerde verdeling van de notenmengeling correspondeert met de verdeling zoals vermeld op de verpakking. Op basis van onze steekproef ($n = 200$) blijkt dat de informatie die op de verpakking vermeld staat misleidend is, $\chi^2(3) = 10.4, p = .015$. De notenmix bevat voornamelijk te veel pinda's en te weinig hazelnoten.

7 Enkelvoudige lineaire regressie

Beschouw de volgende eenvoudige dataset waarbij Y de afhankelijke variabele is en X de (enige) predictor.

Y	X
4	3
4	4
5	5
5	4
6	5
6	6
7	7
7	8

Opgave

- Is X een goede voorspeller voor Y? ($\alpha = 0.05$)
- Hoeveel variantie wordt verklaard door dit model?
- Wat zijn de predicties $\hat{y}_i$ voor elk van de 8 observaties?

Oplossing

Net zoals bij Anova kunnen we twee procedures volgen: ofwel via het lineaire regressie-programma, ofwel via het General Linear Model programma:

- *Linear Regression*

1. **Procedure**

```
Analyze --> Regression --> Linear ...
  Dependent: Y
  Independent(s): X
```

```
[Ok]
```

2. Om de predicties te bekomen:

```
Analyze --> Regression --> Linear ...
  Dependent: Y
  Independent(s): X
  [Save...]
    Predicted Values:
    v Unstandardized
  [Continue]
[Ok]
```

- *GLM*

1. **Procedure**

```
Analyze -> General Linear Model -> Univariate ...
    Dependent Variable: Y
    Covariate(s): X
    [Options...]
        Display
          v Parameter Estimates
        [Continue]
    [Save...]
        Predicted Values
          v Unstandardized
        [Continue]
[Ok]
```

Conclusie

De regressiecoëfficiënt voor X is $b_X = 0.667$. Dit wil zeggen dat indien X stijgt met één eenheid, Y zal stijgen met 0.667. Er is dus een positief verband tussen X en Y. De bekomen overschrijdingskans dienen we te vergelijken met onze vooropgestelde $\alpha = 0.05$.

We verwerpen de nulhypothese daar $p = 0.001 < 0.05$. We dienen onze nulhypothese $(H_0 : \beta_1 = 0)$ te verwerpen.

Ook op basis van de F-toets kunnen we besluiten dat er een significant verband is tussen onze predictor en onze responsvariabele, $F(1, 6) = 39.00, p = .001$. We dienen de nulhypothese $(H_0 : R^2 = 0)$ te verwerpen: de predicties bekomen op basis van het model met de predictor zijn significant beter dan deze bekomen op basis van het nulmodel. $R^2 = 0.867$, wat wil zeggen dat 86.7% van de variantie in Y verklaard wordt door X. Beide toetsen resulteren in het zelfde besluit: er is inderdaad een (positieve) lineaire relatie is tussen X en Y.

Bemerk dat er bij enkelvoudige regressie een relatie is tussen F-statistiek en de t-statistiek:

$$F(1, n-2) = t^2(n-2)$$

Inderdaad: $t^2 = 6.245^2 = 39.00 = F$

Oefening

Lamantijnen zijn een soort rondstaartige zeekoeien die in het ondiepe water voor de kust van Florida leven. Elk jaar worden vele lamantijnen door speedboten gedood of verwond. Hieronder volgen voor de periode van 1977 tot 1990 de gegevens over gedode lamantijnen en aantallen geregistreerde speedboten in Florida (in duizenden boten).

Jaar	Geregistreerde speedboten (X)	Gedode lamantijnen (Y)
1977	447	13
1978	460	21
1979	481	24
1980	498	16
1981	513	24
1982	512	20
1983	526	15
1984	559	34
1985	585	33
1986	614	33
1987	645	39
1988	675	43
1989	711	50
1990	719	47

opgave

- Maak een boxplot voor beide variabelen. Zijn er outliers?
- Hoeveel lamantijnen zijn er gemiddeld gedood tussen 1977 en 1990? Is dit meer dan $\mu = 30$?
- Bereken de correlatie tussen het aantal boten en het aantal gedode lamantijnen.
- Bepaal de determinatiecoëfficiënt voor deze gegevens en ga na of een significant gedeelte van de variantie in de afhankelijke variabele verklaard wordt door de onafhankelijke variabele.
- Is er een sterk bewijs dat het aantal gedode lamantijnen met het aantal speedboten

samenhangt? Formuleer deze vraag als nulhypothese en alternatieve hypothese omtrent de helling (β_1) van de regressielijn, bepaal de t-grootheid en vermeld je conclusie.