Statistiek voor MBW / KW
Uitwerkingen huiswerkopgaven

Dr. ir. D.A.M.P. Blom & Dr. J.B.M. Melissen

2025

Week 12: verschiltoetsen

Hoofdstuk 11

Opdracht 11.m1: Voor een tweetal normale verdelingen is gegeven dat hun standaarddeviatie respectievelijk $\sigma_1 = 4$ en $\sigma_2 = 6$ bedragen. Men wil toetsen $H_0 : \mu_1 = \mu_2$. Een steekproef van negen waarnemingen uit de eerste verdeling leverde een gemiddelde uitkomst op van 48, en een steekproef van zestien waarnemingen uit de tweede verdeling leverde een gemiddelde op van 42. De toetsingsgrootheid bij deze toets levert een z-waarde op van ...

(a) 6.

(b) $2,99$.

(c) $6,63$.

(d) $2,67$.

Uitwerking

We hebben te maken met twee normaal verdeelde kansvariabelen, namelijk $X \sim N(\mu_1 =?; \sigma_1 = 4)$ en $Y \sim N(\mu_2 =?; \sigma_2 = 6)$. Aangezien we respectievelijk steekproeven hebben van groottes $n = 9$ en $m = 16$, geldt voor de theoretische steekproefgemiddeldes:

$$\overline{X} \sim N(\mu = \mu_1; \sigma = \frac{\sigma_1}{\sqrt{n}}) = N(\mu_1 = ?; \sigma = \frac{4}{3})$$
$$\overline{Y} \sim N(\mu = \mu_2; \sigma = \frac{\sigma_2}{\sqrt{m}}) = N(\mu_2 = ?; \sigma = \frac{3}{2})$$

De verschilvariabele $V = \overline{X} - \overline{Y}$ is dan normaal verdeeld met verwachtingswaarde $\mu_V = \mu_1 - \mu_2$ en standaardafwijking $\sigma_V = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} = \sqrt{\frac{16}{9} + \frac{9}{4}} \approx 2,0069$. Onder de nulhypothese geldt $\mu_1 = \mu_2$, oftewel de verschilvariabele heeft verwachtingswaarde $\mu_V = \mu_1 - \mu_2 = 0$. Het geobserveerde verschil is $v = \overline{x} - \overline{y} = 48 - 42 = 6$.
De z-score die hoort bij deze uitkomst is gelijk aan

$$z = \frac{v - \mu_V}{\sigma_V} = \frac{6 - 0}{2,0069} \approx 2,99$$

Het juiste antwoord is dus (b).

Opdracht 11.m3: Voor twee populaties mag worden verondersteld dat deze allebei een normale verdeling volgen en dezelfde onbekende variantie hebben. Men wil toetsen of deze populaties hetzelfde gemiddelde hebben. Daartoe neemt men uit beide populaties een steekproef van tien waarnemingen. Bij de toets is het aantal vrijheidsgraden van de toetsingsgrootheid gelijk aan . . .

(a) 9.

(b) 18.

(c) 19.

(d) – (er is bij deze toetsingsgrootheid geen sprake van vrijheidsgraden).

Uitwerking

We hebben te maken met twee normaal verdeelde kansvariabelen:

$$X_1 \sim N(\mu_1 =?; \sigma_1 =?)$$
$$X_2 \sim N(\mu_2 =?; \sigma_2 =?)$$

Beide kansvariabelen hebben dezelfde onbekende variantie, dat wil zeggen $\sigma = \sigma_1 = \sigma_2$. In dat geval mogen de twee steekproefschattingen s_1 en s_2 voor de standaardafwijking (op basis van steekproeven van grootte $n = m = 10$ respectievelijk) worden gecombineerd tot één schatter (de zogenaamde “pooled variance”).
De kansverdeling die de bijbehorende toetsingsgrootheid volgt is de t-verdeling met $df = n + m - 2 = 10 + 10 - 2 = 18$ vrijheidsgraden. Het juiste antwoord is dus (b).

Opdracht 11.m6: Voor een tweetal normale verdelingen willen we toetsen of ze dezelfde waarde van de variantie hebben. Voor de eerste variabele werd met een steekproef van twaalf waarnemingen de waarde $s^2 = 240$ gevonden. Bij de tweede variabele leverde de steekproef van tien waarnemingen de waarde $s^2 = 160$. De toetsingsgrootheid krijgt hier dan de waarde . . .

(a) 80.

(b) $1,50$.

(c) 4.

(d) $1,25$.

Uitwerking

We hebben te maken met twee normaal verdeelde kansvariabelen

$$X_1 \sim N(\mu_1 =?; \sigma_1 =?)$$
$$X_2 \sim N(\mu_2 =?; \sigma_2 =?).$$

Van deze twee kansvariabelen willen we toetsen of de varianties gelijk zijn, oftewel dat geldt $\sigma^2 = \sigma_1^2 = \sigma_2^2$. Hiervoor moeten we een F-toets uitvoeren.
Voor een F-toets werken we met de volgende nulhypothese H_0 en de alternatieve hypothese H_1.

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ versus } H_1 : \sigma_1^2 \neq \sigma_2^2.$$

Verder hebben we voor beide kansvariabelen twee steekproeven van respectievelijk grootte $n = 12$ en $m = 10$ getrokken om de varianties te schatten, met als resultaat

$s_1^2 = 240$ en $s_2^2 = 160$. De toetsingsgrootheid F is gelijk aan

$$F = \frac{S_1^2}{S_2^2}$$

en volgt onder H_0 de $F(n-1, m-1)$-verdeling. De geobserveerde toetsingsgrootheid is in dit geval gelijk aan

$$f = \frac{s_1^2}{s_2^2} = \frac{240}{160} = 1,50.$$

Het juiste antwoord is dus (b).

Opdracht 11.1: Een instantie die toezicht houdt op de mate van luchtverontreiniging doet regelmatig metingen op diverse plaatsen in het land. De gemeten verontreiniging wordt uitgedrukt in een bepaalde index, die op alle werkdagen wordt bepaald. In 2018 werden enkele nieuwe richtlijnen voor de industrie afgekondigd met als doel het niveau van luchtverontreiniging te verlagen. Om het effect van deze maatregelen te bestuderen, werd een onderzoek gedaan. Hierbij werden 20 waarden (x_i) die bepaald zijn in februari 2017 vergeleken met 20 waarden (y_i) die zijn gemeten in februari 2018. Voor deze waarden werd berekend:

$$\overline{x} = 164,6 \text{ en } s_X = 17,2; \text{ en } \overline{y} = 143,2 \text{ en } s_Y = 15,9$$

Met een verschiltoets gaan we bepalen of er een significante verbetering is opgetreden.

Uitwerking

Laat X en Y de kansvariabelen zijn die het niveau van luchtverontreiniging meet op een willekeurige dag in respectievelijk februari 2017 en februari 2018. Van deze kansvariabelen zijn zowel de verwachtingswaarden als de standaardafwijkingen onbekend. Om te bepalen of er een significante verbetering is opgetreden, moeten we kijken of de gemiddelde luchtverontreiniging is gedaald in 2018. We voeren hiervoor een hypothesetoets uit met de volgende nulhypothese H_0 en de alternatieve hypothese H_1:

$$H_0 : \mu_X \leq \mu_Y \text{ versus } H_1 : \mu_X > \mu_Y.$$

Omdat de standaardafwijkingen onbekend zijn, moeten we werken met een onafhankelijke t-toets met behulp van de verschilvariabele V. Voor deze verschilvariabele V geldt dat

$$\mu_V = \mu_X - \mu_Y$$
$$s_V = \sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}.$$

We veronderstellen dat beide steekproeven te beschouwen zijn als trekking uit (normale) verdelingen met dezelfde variantie, dus kunnen we gebruik maken van de "pooled variance":

$$s_P^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n-1+m-1} = \frac{19 \cdot 17,2 + 19 \cdot 15,9}{38} = 16,55.$$

Onder de nulhypothese H_0 (uitgaande van het extreme geval $\mu_X = \mu_Y$) geldt dus voor de verschilvariabele V dat

$$\mu_V = \mu_X - \mu_Y = 0$$
$$s_V = \sqrt{\frac{s_P^2}{n} + \frac{s_P^2}{m}} = \sqrt{\frac{16,55}{20} + \frac{16,55}{20}} \approx 1,2865$$

De t-waarde bijbehorende bij de uitkomst $v = \overline{x} - \overline{y} = 164,6 - 143,2 = 21,4$ is gelijk aan

$$t = \frac{v - \mu_V}{s_V} = \frac{21,4 - 0}{1,2865} = 16,6347.$$

Merk op dat deze t-waarde uit een t-verdeling komt met df $= n - m + 2 = 38$ vrijheidsgraden.
Aangezien we rechtszijdig toetsen, kunnen we de p-waarde als volgt berekenen:

$$p = \text{tcdf}(\text{lower} = 16,63; \text{upper} = 10^{99}; \text{df} = 38) \approx 2,4554 \cdot 10^{-19}.$$

De p-waarde is extreem klein (kleiner dan welke vaakgebruikte keuze voor α dan ook), dus verwerpen we de nulhypothese. Er is voldoende reden om op basis van deze steekproef aan te nemen dat er daadwerkelijk een significante daling is gerealiseerd in de luchtverontreiniging in 2018.

Opdracht 11.3: De variabele X is normaal verdeeld met onbekende μ en een variantie Var$(X) = 100$. Ook de variabele Y is normaal verdeeld met onbekende verwachtingswaarde. Er geldt Var$(Y) = 30$. Van de variabele X worden tien trekkingen gedaan. Die leveren als gemiddelde $\overline{x} = 50$. Er worden vijf trekkingen gedaan van de variabele Y die als gemiddelde opleveren $\overline{y} = 55$. Toets of beide variabelen dezelfde verwachtingswaarde kunnen hebben (kies $\alpha = 0,05$).

Uitwerking

We hebben te maken met twee normaal verdeelde kansvariabelen

$$X \sim N(\mu_X = ?; \sigma_X = \sqrt{100} = 10)$$
$$Y \sim N(\mu_Y = ?; \sigma_Y = \sqrt{30}).$$

Om te toetsen of de gemiddeldes van beide kansvariabelen gelijk zijn, gebruiken we de volgende nulhypothese H_0 en alternatieve hypothese H_1:

$$H_0 : \mu_X = \mu_Y \text{ versus } H_1 : \mu_X \neq \mu_Y.$$

Verder zijn het significantieniveau $\alpha = 0,05$ en de steekproefgemiddeldes $\overline{x} = 50$ en $\overline{y} = 55$ gegeven. Als toetsingsgrootheid kijken we naar de verschilvariabele $V = \overline{X} - \overline{Y}$. Hiervoor geldt dat

$$V = \overline{X} - \overline{Y} = N\left(\mu_V = \mu_X - \mu_Y; \sigma_V = \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} = \sqrt{\frac{100}{10} + \frac{30}{5}} = 4\right).$$

Onder de nulhypothese geldt $\mu_X = \mu_Y$, oftewel de verschilvariabele heeft verwachtingswaarde $\mu_V = \mu_X - \mu_Y = 0$.
Methode 1 (kritiek gebied):
Omdat we tweezijdig toetsen, is het kritieke gebied van de vorm $(-\infty; g_1]$ en $[g_2; \infty)$, waarbij de grenzen bepaald kunnen worden met:

$$g_1 = \text{InvNorm}(\text{opp} = \alpha/2 = 0,025; \mu = \mu_V = 0; \sigma = \sigma_V = 4) \approx -7,8399$$
$$g_2 = \text{InvNorm}(\text{opp} = 1 - \alpha/2 = 0,975; \mu = \mu_V = 0; \sigma = \sigma_V = 4) \approx 7,8399$$

Het geobserveerde verschil is $v = \overline{x} - \overline{y} = 50 - 55 = -5$. Deze waarde ligt niet in het kritieke gebied, dus we accepteren de nulhypothese H_0. Er is onvoldoende bewijs om de aanname te verwerpen dat beide variabelen dezelfde verwachtingswaarde zouden hebben.
Methode 2 (p-waarde):
Het geobserveerde verschil is $v = \overline{x} - \overline{y} = 50 - 55 = -5$ en ligt dus links van het gemiddelde $\mu_V = 0$. Omdat we tweezijdig toetsen, berekenen we de p-waarde als de linkeroverschrijdingskans van de waarde -5 en vergelijken deze met $\alpha/2 = 0,025$.

$$p = \text{normalcdf}(\text{lower} = -10^{99}; \text{upper} = -5; \mu = 0; \sigma = 4) = 0,1056$$

De p-waarde is groter dan $\alpha/2$, dus we accepteren de nulhypothese H_0. Er is onvoldoende bewijs om de aanname te verwerpen dat beide variabelen dezelfde verwachtingswaarde zouden hebben.

Opdracht 11.5: Van werknemers in een bepaalde bedrijfstak wordt het verband tussen jaarsalaris en het bezit van een vakdiploma onderzocht. Groep 1 is een steekproef uit de populatie van werknemers met diploma, groep 2 is een steekproef van werknemers zonder diploma. De resultaten waren (jaarinkomens in duizenden euro's):

Groep 1	40	45	48	33	42	35	32	47	38	37	43
Groep 2	25	28	30	35	38	32	22				

Veronderstel dat beide inkomensverdelingen mogen worden beschouwd als normale verdelingen met dezelfde onbekende variantie. Toets of $\mu_X = \mu_Y$ (kies $\alpha = 0,05$).

Uitwerking

We hebben te maken met twee normaal verdeelde kansvariabelen

$$X \sim N(\mu_X =?; \sigma_X =?)$$
$$Y \sim N(\mu_Y =?; \sigma_Y =?).$$

Om te toetsen of de gemiddeldes van beide kansvariabelen gelijk zijn, gebruiken we de volgende nulhypothese H_0 en alternatieve hypothese H_1:

$$H_0 : \mu_X = \mu_Y \text{ versus } H_1 : \mu_X \neq \mu_Y.$$

Verder is het significantieniveau $\alpha = 0,05$ gegeven.
In de tabel zijn data geven voor beide groepen. We berekenen allereerst de steekproefgemiddeldes:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{40 + 45 + \ldots + 43}{12} = 40$$
$$\overline{y} = \frac{y_1 + y_2 + \ldots + y_m}{m} = \frac{25 + 28 + \ldots + 22}{7} = 30$$

Met behulp van de steekproefgemiddeldes kunnen we nu de steekproefvarianties bepalen.

$$s_X^2 = \frac{(x_1-\overline{x})^2+\ldots+(x_n-\overline{x})^2}{n-1} = \frac{(40-40)^2+\ldots+(43-40)^2}{11} = 30,2$$
$$s_Y^2 = \frac{(y_1-\overline{y})^2+\ldots+(y_m-\overline{y})^2}{m-1} = \frac{(25-30)^2+\ldots+(22-30)^2}{6} = 31$$

Als toetsingsgrootheid kijken we naar de verschilvariabele $V=\overline{X}-\overline{Y}$. Deze verschilvariabele heeft een verwachtingswaarde $\mu_V=\mu_X-\mu_Y$. Omdat we mogen uitgaan van gelijke varianties, kunnen we de standaardafwijking σ_V van V schatten met behulp van de "pooled variance":

$$s_P^2 = \frac{(n-1)s_X^2+(m-1)s_Y^2}{n-1+m-1} = \frac{11\cdot 30,2+6\cdot 31}{17} = 30,4824.$$

Een schatting van σ_V berekenen we dan als volgt:

$$s_V = \sqrt{\frac{s_P^2}{n}+\frac{s_P^2}{m}} = \sqrt{\frac{30,4824}{12}+\frac{30,4824}{7}} \approx 2,6258$$

Het geobserveerde verschil is $v=\overline{x}-\overline{y}=40-30=10$. Onder de nulhypothese geldt $\mu_X=\mu_Y$, oftewel de verschilvariabele heeft verwachtingswaarde $\mu_V=\mu_X-\mu_Y=0$. De bijbehorende t-waarde berekenen we dan als volgt:

$$t=\frac{v-\mu_V}{s_V}=\frac{10-0}{2,6258}\approx 3,8084.$$

Methode 1 (kritiek gebied):
Omdat we tweezijdig toetsen, is het kritieke gebied van de vorm $(-\infty;g_1]$ en $[g_2;\infty)$, waarbij de grenzen bepaald kunnen worden met de t-verdeling. Omdat we met de pooled variance werken, is het aantal vrijheidsgraden gelijk aan $\text{df}=n+m-2=17$.

$$g_1 = \text{InvT}(\text{opp}=\alpha/2=0,025;\text{df}=17)\approx -2,1098$$
$$g_2 = \text{InvT}(\text{opp}=1-\alpha/2=0,975;\text{df}=17)\approx 2,1098$$

De berekende $t\approx 3,8084$ ligt dus in het kritieke gebied, dus we verwerpen de nulhypothese H_0. Er is op basis van deze steekproeven voldoende reden om aan te nemen dat het gemiddelde jaarsalaris significant verschilt tussen beide groepen werknemers.
Methode 2 (p-waarde):
De berekende $t\approx 3,8084$ ligt rechts van het gemiddelde $\mu_V=0$. Omdat we tweezijdig toetsen, berekenen we de p-waarde als de rechteroverschrijdingskans van de waarde $3,8084$ en vergelijken deze met $\alpha/2=0,025$.

$$p=\text{normalcdf}(\text{lower}=3,8084;\text{upper}=10^{99};\text{df}=17)\approx 7,0254\cdot 10^{-4}.$$

De p-waarde is (veel) kleiner dan $\alpha/2$, dus we verwerpen de nulhypothese H_0. Er is op basis van deze steekproeven voldoende reden om aan te nemen dat het gemiddelde jaarsalaris significant verschilt tussen beide groepen werknemers.

Opdracht 11.7: In een studie naar luchtvervuiling leverde een aselecte steekproef van acht luchtmonsters in de omgeving van een bepaalde fabriek gemiddeld $2,26$ microgram van een schadelijke stof per m^3, met een standaardafwijking van $0,56$. In een grote stad leverde een steekproef van tien monsters een resultaat van $1,54$ als gemiddelde en $0,42$ als standaardafwijking. Ook in dit geval ging het om de hoeveelheid van dezelfde stof per m^3. Onderzoek met $\alpha = 0,01$ of het gevonden verschil significant is. Hierbij mag men aannemen dat de steekproeven afkomstig waren uit twee normale verdelingen met dezelfde variantie.

Uitwerking

We hebben te maken met twee normaal verdeelde kansvariabelen die het aantal microgram schadelijke stof per m^3 meet, respectievelijk in de omgeving van een bepaalde fabriek en in een grote stad.

$$X \sim N(\mu_X =?; \sigma_X =?)$$
$$Y \sim N(\mu_Y =?; \sigma_Y =?).$$

Om te toetsen of de gemiddeldes van beide kansvariabelen gelijk zijn, gebruiken we de volgende nulhypothese H_0 en alternatieve hypothese H_1:

$$H_0 : \mu_X = \mu_Y \text{ versus } H_1 : \mu_X > \mu_Y \text{ (hogere luchtvervuiling rondom fabriek)}.$$

Verder is het significantieniveau $\alpha = 0,01$ gegeven. De steekproefdata uit de vraag kunnen als volgt worden samengevat:

$$\begin{array}{lll} \overline{x} = 2,26 & s_X = 0,56 & n = 8 \\ \overline{y} = 1,54 & s_Y = 0,42 & m = 10 \end{array}$$

Als toetsingsgrootheid kijken we naar de verschilvariabele $V = \overline{X} - \overline{Y}$. Deze verschilvariabele heeft een verwachtingswaarde $\mu_V = \mu_X - \mu_Y$. Omdat we mogen uitgaan van gelijke varianties, kunnen we de standaardafwijking σ_V van V schatten met behulp van de “pooled variance”:

$$s_P^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n-1+m-1} = \frac{7 \cdot 0,56^2 + 9 \cdot 0,42^2}{16} \approx 0,2364.$$

Een schatting van σ_V berekenen we dan als volgt:

$$s_V = \sqrt{\frac{s_P^2}{n} + \frac{s_P^2}{m}} = \sqrt{\frac{0,2364}{8} + \frac{0,2364}{10}} \approx 0,2306.$$

Het geobserveerde verschil is $v = \overline{x} - \overline{y} = 2,26 - 1,54 = 0,72$. Onder de nulhypothese geldt $\mu_1 = \mu_2$, oftewel de verschilvariabele heeft verwachtingswaarde $\mu_V = \mu_1 - \mu_2 = 0$. De bijbehorende t-waarde berekenen we dan als volgt:

$$t = \frac{v - \mu_V}{s_V} = \frac{0,72 - 0}{0,2306} \approx 3,1217.$$

Methode 1 (kritiek gebied):
Omdat we rechtszijdig toetsen, is het kritieke gebied van de vorm $[g; \infty)$, waarbij de grens g bepaald kan worden met de t-verdeling. Omdat we met de pooled variance werken, is het aantal vrijheidsgraden gelijk aan df$= n + m - 2 = 16$.

$$g = \mathrm{InvT}(\text{opp} = 1 - \alpha = 0,99; \text{df} = 16) \approx 2,5835$$

De berekende $t \approx 3,1217$ ligt dus in het kritieke gebied, dus we verwerpen de nulhypothese H_0. Er is op basis van deze steekproeven voldoende reden om aan te nemen dat de gemiddelde luchtvervuiling significant hoger is in de omgeving van de fabriek.
Methode 2 (p-waarde):
Omdat we rechtszijdig toetsen, berekenen we de p-waarde als de rechteroverschrijdingskans van de waarde $3,1217$ en vergelijken deze met $\alpha = 0,01$.

$$p = \text{normalcdf}(\textsf{lower} = 3,1217; \textsf{upper} = 10^{99}; \textsf{df} = 16) \approx 0,0033.$$

De p-waarde is kleiner dan het significantieniveau α, dus we verwerpen de nulhypothese H_0. Er is op basis van deze steekproeven voldoende reden om aan te nemen dat de gemiddelde luchtvervuiling significant hoger is in de omgeving van de fabriek.

Opdracht 11.12: Een grootwinkelbedrijf verkoopt in alle filialen een bepaald afwasmiddel. In een week werd in zes filialen het volgende aantal flessen van dit afwasmiddel verkocht.

Filiaal	A	B	C	D	E	F
Aantal flessen	320	260	440	400	380	300

In de week nadat voorgaande gegevens werden verzameld, is een intensieve reclamecampagne gehouden voor het desbetreffende afwasmiddel. Volgend op deze reclamecampagne worden in tien filialen in één week de volgende aantallen flessen verkocht:

Filiaal	P	Q	R	S	T	U	V	W	X	Y
Aantal flessen	400	360	410	500	540	380	490	520	420	480

Toets of het gemiddelde aantal verkochte flessen voor en na de behandeling gelijk is ($\alpha = 0,05$). (Ga ervan uit dat de waargenomen aantallen mogen worden beschouwd als trekkingen uit normale verdelingen.)

Uitwerking

Laat X en Y de kansvariabelen zijn die het aantal verkochte flessen afwasmiddel telt respectievelijk vóór en na de intensieve reclamecampagne. We mogen ervan uitgaan dat deze aantallen trekkingen zijn uit normale verdelingen, oftewel

$$X \sim N(\mu_X =?; \sigma_X =?)$$
$$Y \sim N(\mu_Y =?; \sigma_Y =?).$$

Merk op dat we niet weten of de varianties gelijk zijn, dus hiervoor moeten we eerst een F-toets uitvoeren.
***F*-toets**
We gaan toetsen of de varianties van X en Y gelijk zijn. Hiervoor definiëren we allereerst de nulhypothese H_0 en alternatieve hypothese H_1:

$$H_0 : \sigma_X^2 = \sigma_Y^2 \text{ versus } H_1 : \sigma_X^2 \neq \sigma_Y^2.$$

We gaan opnieuw uit van een significantieniveau $\alpha = 0,05$ en de steekproefgegevens in de bovenstaande tabellen. We berekenen allereerst de steekproefgemiddeldes:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{320 + 260 + \ldots + 300}{6} = 350$$
$$\overline{y} = \frac{y_1 + y_2 + \ldots + y_m}{m} = \frac{400 + 360 + \ldots + 480}{10} = 450$$

Met behulp van de steekproefgemiddeldes kunnen we nu de steekproefvarianties bepalen.

$$s_X^2 = \frac{(x_1 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1} = \frac{(320-350)^2 + \ldots + (300-350)^2}{5} = 4600$$
$$s_Y^2 = \frac{(y_1 - \overline{y})^2 + \ldots + (y_m - \overline{y})^2}{m-1} = \frac{(400-450)^2 + \ldots + (480-450)^2}{9} = 4000$$

De toetsingsgrootheid van de F-toets is gelijk aan

$$F = \frac{S_X^2}{S_Y^2}$$

en volgt de F-verdeling met respectievelijk df1 $= n - 1 = 5$ en df2 $= m - 1 = 9$ vrijheidsgraden. In dit geval hebben we een toetsingsgrootheid van

$$f = \frac{s_X^2}{s_Y^2} = \frac{4600}{4000} = 1,15$$

Methode 1 (kritiek gebied):
Omdat we tweezijdig toetsen, is het kritieke gebied van de vorm $(-\infty; g_1]$ en $[g_2; \infty)$, waarbij de grenzen g_1 en g_2 bepaald kunnen worden met de $F(5,9)$-verdeling.

$$\mathrm{Fcdf}(\text{lower} = 0; \text{upper} = g_1; \text{df1} = 5; \text{df2} = 9) = \alpha/2 = 0,025 \to g_1 \approx 0,1497$$
$$\mathrm{Fcdf}(\text{lower} = g_2; \text{upper} = 10^{99}; \text{df1} = 5; \text{df2} = 9) = \alpha/2 = 0,025 \to g_2 \approx 4,4844$$

De berekende $f = 1,15$ ligt dus niet in het kritieke gebied, dus we accepteren de nulhypothese H_0. Er is op basis van deze steekproeven onvoldoende reden om aan te nemen dat de varianties niet gelijk zouden zijn aan elkaar.
Methode 2 (p-waarde):
De berekende $f = 1,15$ ligt boven de 1, dus we berekenen de p-waarde als de rechteroverschrijdingskans van de waarde $1,15$ en vergelijken deze met $\alpha/2 = 0,025$.

$$p = \mathrm{normalcdf}(\text{lower} = 1,15; \text{upper} = 10^{99}; \text{df1} = 5; \text{df2} = 9) \approx 0,4020.$$

De p-waarde is groter dan het significantieniveau α, dus we accepteren de nulhypothese H_0. Er is op basis van deze steekproeven onvoldoende reden om aan te nemen dat de varianties niet gelijk zouden zijn aan elkaar.
Onafhankelijke *t*-toets (gelijke variantie)
Door het toetsresultaat van de F-toets mogen we uitgaande van gelijke varianties, oftewel $\sigma^2 = \sigma_X^2 = \sigma_Y^2$. We gaan toetsen of de gemiddeldes van X en Y gelijk zijn. Hiervoor definiëren we allereerst de nulhypothese H_0 en alternatieve hypothese H_1:

$$H_0 : \mu_X = \mu_Y \text{ versus } H_1 : \mu_X = \mu_Y.$$

Gegeven is het significantieniveau $\alpha = 0,05$ en de steekproefgegevens hierboven berekend:

$$\overline{x} = 350 \quad s_X^2 = 4600 \quad n = 6$$
$$\overline{y} = 450 \quad s_Y^2 = 4000 \quad m = 10$$

Als toetsingsgrootheid kijken we naar de verschilvariabele $V = \overline{X} - \overline{Y}$. Deze verschilvariabele heeft een verwachtingswaarde $\mu_V = \mu_X - \mu_Y$. Omdat we mogen uitgaan van gelijke varianties, kunnen we de standaardafwijking σ_V van V schatten met behulp van de "pooled variance":

$$s_P^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n-1+m-1} = \frac{6 \cdot 4600 + 10 \cdot 4000}{14} \approx 4828,5714$$

Een schatting van σ_V berekenen we dan als volgt:

$$s_V = \sqrt{\frac{s_P^2}{n} + \frac{s_P^2}{m}} = \sqrt{\frac{4828,5714}{6} + \frac{4828,5714}{10}} \approx 35,8834.$$

Het geobserveerde verschil is $v = \overline{x} - \overline{y} = 350 - 450 = -100$. Onder de nulhypothese geldt $\mu_1 = \mu_2$, oftewel de verschilvariabele heeft verwachtingswaarde $\mu_V = \mu_1 - \mu_2 = 0$. De bijbehorende t-waarde berekenen we dan als volgt:

$$t = \frac{v - \mu_V}{s_V} = \frac{-100 - 0}{35,8834} \approx -2,7868.$$

Methode 1 (kritiek gebied):
Omdat we tweezijdig toetsen, is het kritieke gebied van de vorm $(-\infty; g_1]$ en $[g_2; \infty)$, waarbij de grenzen g_1 en g_2 bepaald kunnen worden met de t-verdeling. Omdat we met de pooled variance werken, is het aantal vrijheidsgraden gelijk aan df $= n+m-2 = 14$.

$$g_1 = \mathrm{InvT}(\text{opp} = \alpha/2 = 0,025; \text{df} = 14) \approx -2,1448$$
$$g_2 = \mathrm{InvT}(\text{opp} = 1 - \alpha/2 = 0,975; \text{df} = 14) \approx 2,1448$$

De berekende $t \approx -2,7868$ ligt dus in het kritieke gebied, dus we verwerpen de nulhypothese H_0. Er is op basis van deze steekproeven voldoende reden om aan te nemen dat het aantal verkochte flessen afwasmiddel significant hoger is na de intensieve reclamecampagne.

Methode 2 (p-waarde):
De berekende $t \approx -2,7868$ ligt links van het gemiddelde $\mu_V = 0$. Omdat we tweezijdig toetsen, berekenen we de p-waarde als de linkeroverschrijdingskans van de waarde $-2,7868$ en vergelijken deze met $\alpha/2 = 0,025$.

$$p = \mathrm{tcdf}(\text{lower} = -10^{99}; \text{upper} = -2,7868; \text{df} = 14) \approx 0,0073.$$

De p-waarde is kleiner dan $\alpha/2$, dus we verwerpen de nulhypothese H_0. Er is op basis van deze steekproeven voldoende reden om aan te nemen dat het aantal verkochte flessen afwasmiddel significant hoger is na de intensieve reclamecampagne.

Opdracht 11.13: In zes filialen van een supermarktketen is het aantal verkochte flessen van een wasmiddel vlak voor en vlak na een reclamecampagne geregistreerd. De resultaten waren:

Filiaal	**A**	**B**	**C**	**D**	**E**	**F**
Aantal voor	320	260	440	400	380	310
Aantal na	420	380	520	490	490	400

Toets of er een opvallend verschil is in de omzet per filiaal voor en na de campagne ($\alpha = 0,05$). Voer de toets op twee manieren uit en geef aan van welke veronderstellingen men is uitgegaan bij de toetsen (vergelijk opgave 11.12).

Uitwerking

Laat X en Y de kansvariabelen zijn die het aantal verkochte flessen wasmiddel telt respectievelijk vóór en na de reclamecampagne. We kunnen ervan uit dat deze aantallen *gepaard* zijn

$$X \sim N(\mu_X =?; \sigma_X =?)$$
$$Y \sim N(\mu_Y =?; \sigma_Y =?).$$

Hierdoor kunnen we de gepaarde uitkomsten samennemen tot een nieuwe variabele gebaseerd op het verschil:

Filiaal	**A**	**B**	**C**	**D**	**E**	**F**
Aantal voor (x)	320	260	440	400	380	310
Aantal na (y)	420	380	520	490	490	400
Verschil ($v = y - x$)	100	120	80	90	110	90

We gaan toetsen of het gemiddelde van V gelijk is aan 0, omdat dit zou duiden op $\mu_X = \mu_Y$. Hiervoor definiëren we allereerst de nulhypothese H_0 en alternatieve hypothese H_1:

$$H_0 : \mu_V = 0 \text{ versus } H_1 : \mu_V > 0 \text{(significant hogere aantallen na reclamecampagne)}.$$

Gegeven is het significantieniveau $\alpha = 0,05$ en de steekproefgegevens in bovenstaande tabel. Het steekproefgemiddelde van de verschillen is gelijk aan

$$\overline{v} = \frac{v_1 + v_2 + \ldots + v_6}{6} = \frac{100 + 120 + \ldots + 90}{6} \approx 98,3333$$

De steekproefvariantie s_V^2 berekenen we als volgt:

$$s_V^2 = \frac{(v_1 - \overline{v})^2 + \ldots + (v_n - \overline{v})^2}{n-1} = \frac{(100 - 98,3333)^2 + \ldots + (90 - 98,3333)^2}{5} = 216,6667$$

Neem aan dat V normaal verdeeld is $\mu_V = 0$ en met onbekende standaardafwijking σ_v. Hierdoor moeten we gebruik maken van de t-verdeling met df $= n - 1 = 5$ vrijheidsgraden. Onder de nulhypothese is de toetsingsgrootheid van deze toets gelijk aan

$$t = \frac{\overline{v} - \mu_V}{s_V} = \frac{98,3333 - 0}{\sqrt{216,6667}} \approx 6,6804.$$

Methode 1 (kritiek gebied):

Omdat we rechtszijdig toetsen, is het kritieke gebied van de vorm $[g;\infty)$, waarbij de grens g bepaald kan worden met de t-verdeling.

$$g_2 = \text{InvT}(\text{opp} = 1-\alpha = 0,95; \text{df} = 5) \approx 2,0150$$

De berekende $t \approx 6,6804$ ligt dus in het kritieke gebied, dus we verwerpen de nulhypothese H_0. Er is op basis van deze steekproeven voldoende reden om aan te nemen dat het aantal verkochte flessen afwasmiddel significant hoger is na de reclamecampagne.
Methode 2 (p-waarde):
Omdat we rechtszijdig toetsen, berekenen we de p-waarde als de rechteroverschrijdingskans van de waarde $6,6804$ en vergelijken deze met $\alpha/2 = 0,025$.

$$p = \text{tcdf}(\text{lower} = 6,6804; \text{upper} = 10^{99}; \text{df} = 5) \approx 5,6789 \cdot 10^{-4}.$$

De p-waarde is (veel) kleiner dan het significantieniveau α, dus we verwerpen de nulhypothese H_0. Er is op basis van deze steekproeven voldoende reden om aan te nemen dat het aantal verkochte flessen afwasmiddel significant hoger is na de reclamecampagne.

Opdracht 11.14: Een onderzoeker merkt naar aanleiding van de aanpak bij opgave 11.13 op dat het misschien beter is om per filiaal eerst de procentuele omzetstijging te berekenen. Voer op basis van deze uitkomsten nogmaals een t-toets uit. (Kies $\alpha = 0,05$.)

Uitwerking

Hierdoor kunnen we de gepaarde uitkomsten samennemen tot een nieuwe variabele gebaseerd op het verschil:

Filiaal	**A**	**B**	**C**	**D**	**E**	**F**
Aantal voor (x)	320	260	440	400	380	310
Aantal na (y)	420	380	520	490	490	400
Verschil ($v = y - x$)	100	120	80	90	110	90
Verschil % ($pv = \frac{v}{x} \cdot 100\%$)	$31,25$	$46,1538$	$18,1818$	$22,5$	$28,9474$	$29,0323$

We gaan toetsen of het gemiddelde van PV gelijk is aan 0, omdat dit zou duiden op $\mu_X = \mu_Y$. Hiervoor definiëren we allereerst de nulhypothese H_0 en alternatieve hypothese H_1:

$$H_0 : \mu_{PV} = 0 \text{ versus } H_1 : \mu_{PV} > 0 \text{(significant hogere aantallen na reclamecampagne)}.$$

Gegeven is het significantieniveau $\alpha = 0,05$ en de steekproefgegevens in bovenstaande tabel. Het steekproefgemiddelde van de verschillen is gelijk aan

$$\overline{pv} = \frac{pv_1 + pv_2 + \ldots + pv_6}{6} = \frac{31,25 + 46,1538 + \ldots + 29,0323}{6} \approx 29,3442$$

De steekproefvariantie s^2_{PV} berekenen we als volgt:

$$\begin{aligned} s^2_{PV} &= \frac{(v_1 - \overline{v})^2 + \ldots + (v_n - \overline{v})^2}{n-1} \\ &= \frac{(31,25 - 29,3442)^2 + \ldots + (29,0323 - 29,3442)^2}{5} \\ &\approx 91,5786 \end{aligned}$$

Neem aan dat PV normaal verdeeld is $\mu_{PV} = 0$ en met onbekende standaardafwijking σ_{PV}. Hierdoor moeten we gebruik maken van de t-verdeling met df $= n - 1 = 5$ vrijheidsgraden. Onder de nulhypothese is de toetsingsgrootheid van deze toets gelijk aan

$$t = \frac{\overline{pv} - \mu_{PV}}{s_{PV}} = \frac{29,3442 - 0}{\sqrt{91,5786}} \approx 3,0664$$

Methode 1 (kritiek gebied):
Omdat we rechtszijdig toetsen, is het kritieke gebied van de vorm $[g; \infty)$, waarbij de grens g bepaald kan worden met de t-verdeling.

$$g = \mathrm{InvT}(\text{opp} = 1 - \alpha = 0,95; \text{df} = 5) \approx 2,0150$$

De berekende $t \approx 3,0664$ ligt dus in het kritieke gebied, dus we verwerpen de nulhypothese H_0. Er is op basis van deze steekproeven voldoende reden om aan te nemen dat het aantal verkochte flessen afwasmiddel significant hoger is na de reclamecampagne.
Methode 2 (p-waarde):
Omdat we rechtszijdig toetsen, berekenen we de p-waarde als de rechteroverschrijdingskans van de waarde $6,6804$ en vergelijken deze met $\alpha/2 = 0,025$.

$$p = \mathrm{tcdf}(\text{lower} = 3,0664; \text{upper} = 10^{99}; \text{df} = 5) \approx 0,0139.$$

De p-waarde is kleiner dan het significantieniveau α, dus we verwerpen de nulhypothese H_0. Er is op basis van deze steekproeven voldoende reden om aan te nemen dat het aantal verkochte flessen afwasmiddel significant hoger is na de reclamecampagne.