Computerzitting 2:
Hypothesetoetsen in R
2022-2023

Om een bepaalde hypothese te toetsen volgen we steeds dezelfde werkwijze:

- Formuleer de nulhypothese H_0 en de alternatieve hypothese H_1.
- Geef de teststatistiek.
- Controleer de voorwaarden die nodig zijn om de test uit te voeren. Indien nodig, herformuleer de hypothese (bv. niet-parametrisch, transformatie, ...)
- Voer de test uit die bij de hypothese hoort en geef de waarde van de teststatistiek en de bijbehorende p-waarde.
- Formuleer een besluit op basis van de volgende regel:
  - als de p-waarde $\geq \alpha \rightarrow$ verwerp H_0 niet
  - als de p-waarde $< \alpha \rightarrow$ verwerp H_0

  met α het zelf gekozen of opgegeven significantieniveau.

1 Toetsen op het gemiddelde van 1 variabele

1.1 Met behulp van de p-waarde

We willen een test uitvoeren omtrent het ongekend gemiddelde μ van een variabele X die normaal verdeeld is. We veronderstellen σ^2 ongekend. We kunnen 3 gevallen onderscheiden:

1. Voor een **tweezijdige** test is de hypothese:

   $$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu \neq \mu_0$$

   met $\mu_0 \in \mathbb{R}$.

2. Als men wil testen of μ **groter** is dan een bepaald getal μ_0, wordt de hypothese als volgt geformuleerd:

   $$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu > \mu_0$$

   en gaat men een **rechtseenzijdige** test uitvoeren.

3. In het geval men wil testen of μ **kleiner** is dan een bepaald getal μ_0, krijgen we een **linkseenzijdige** test voor de hypothese

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu < \mu_0.$$

Op voorwaarde dat de gegevens normaal verdeeld zijn, kunnen we een hypothese van deze vorm testen met behulp van de volgende teststatistiek:

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \text{ met } S \text{ de steekproefstandaarddeviatie.}$$

Onder H_0 geldt dat T een Student t-verdeling heeft met $n-1$ vrijheidsgraden, men noteert:

$$T \sim t_{n-1}.$$

Met de beschikbare gegevens berekenen we $t = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$. De p-waarde $(= p)$ wordt dan als volgt berekend:

1. voor een tweezijdige test:
   $p = 2P(T < t)$ als $t < 0$;
   $p = 2P(T > t)$ als $t > 0$
   of algemeen: $p = 2P(T > |t|)$.
2. voor een rechtseenzijdige test: $p = P(T > t)$.
3. voor een linkseenzijdige test: $p = P(T < t)$.

1.2 Toepassing

Op het weerbericht beweert men dat de gemiddelde junitemperatuur op het middaguur in Leuven 20 graden Celsius is. Jij trekt die stelling in twijfel. Je zoekt de junitemperaturen van het jaar 2001 op en je test of de mensen van het KMI gelijk hebben op significantieniveau $\alpha = 0.05$.

De gegevens zijn terug te vinden in de dataset *datajuni.xls* op Blackboard.
Vooraleer de hypothesetest uit te voeren moet je de normaliteit van de variabele temperatuur nagaan. Grafisch kan dit aan de hand van een boxplot, histogram en QQ-plot. Later in deze computerzitting zullen we zien dat we normaliteit van een variabele ook op een formele manier kunnen (en moeten) testen.

Oefening 1

- Wat is de nulhypothese en de alternatieve hypothese?

  .....
  .....

- Welke teststatistiek gebruik je? Wat is de verdeling van de teststatistiek onder H_0?
  .....
  .....

- Zijn de gegevens normaal verdeeld? Ga dit na op een grafische manier.

.....
.....
.....
.....
.....
.....

Als blijkt dat de gegevens normaal verdeeld zijn, kunnen we de hypothese gaan toetsen. In R gebeurt dit op de volgende manier:

```
t.test(datajuni$TEMPERATUUR,mu=20,alternative="two.sided")
```

Hierbij kies je zelf de waarde voor μ_0. In de uitvoer kan je de waarde van de teststatistiek, 95% betrouwbaarheidsinterval en de p-waarde terugvinden.

Opgelet: R geeft automatisch de p-waarde voor de situatie van een tweezijdige test als je *alternative* niet specifieert. Wil je de p-waarde voor een rechtseenzijdige test, dan geef je

```
alternative="greater"
```

in en voor een linkseenzijdige test

```
alternative="less".
```

Opmerking: Extra informatie over de t-test (gebruik en syntax) in R kan je vinden door `help(t.test)` in te geven.

Oefening 2 Teken de p-waarde.
.....
.....
.....
.....
.....

1.3 Met behulp van een betrouwbaarheidsinterval

Er is een verband tussen een **tweezijdige** hypothesetest en betrouwbaarheidsinterval. Het opstellen van een betrouwbaarheidsinterval voor het ongekend gemiddelde μ vormt een tweede mogelijkheid om een tweezijdige hypothese te toetsen.

De experimentele waarde $t = \dfrac{(\bar{x} - 20)}{s/\sqrt{n}}$ ligt in het aanvaardingsgebied $(-t_{n-1;\alpha/2}, t_{n-1;\alpha/2})$ (wat overeenkomt met: de p-waarde $> \alpha$ of verwerp H_0 niet) als en slechts als de waarde μ_0 in het $(100 - \alpha)\%$-betrouwbaarheidsinterval voor μ ligt.

R geeft automatisch bij de t-test een 95%-betrouwbaarheidsinterval. Wil je echter een 99%-betrouwbaarheidsinterval, dan geef je volgende code in R in:

```
t.test(datajuni$TEMPERATUUR,mu=20,conf.level=0.99).
```

Oefening 3 We keren terug naar de hypothese geformuleerd in Oefening 1.

- Wat is het 95% B.I. voor μ?
  .....
  .....
- Ligt $\mu_0 = 20$ in dit interval?

  .....
  .....
- Wat besluit je?

  .....
  .....
- Doe dezelfde test op significantieniveau $\alpha = 0.01$.
  .....
  .....
  .....
  .....

2 Testen op normaliteit: Shapiro-Wilk

Naast grafische methoden om normaliteit na te gaan (een histogram, een boxplot of een normale QQ-plot), bestaan er ook meer formele statistische testen. Eén daarvan is de Shapiro-Wilk test. Om te besluiten of de data al dan niet uit een normale verdeling komen, testen we de volgende hypothese:

$$H_0 : \text{de gegevens komen uit een normale verdeling}$$
$$\text{versus}$$
$$H_1 : \text{de gegevens komen niet uit een normale verdeling.}$$

Deze test berekent een teststatistiek welke overeenkomt met de correlatiecoëfficiënt r_Q van de punten op een normale QQ-plot en ligt dus tussen 0 en 1. Als r_Q dicht bij 1 ligt, is er een sterk lineair verband in de QQ-plot en komen de datapunten waarschijnlijk uit een normale verdeling. Als r_Q niet dicht bij 1 ligt is er een minder goed lineair verband in de QQ-plot. Bijgevolg komen de data dan waarschijnlijk niet uit een normale verdeling. De precieze verdeling van de teststatistiek is niet gekend maar R berekent een benaderende p-waarde die ook in tabellen te vinden is. Als deze p-waarde kleiner is dan 0.05, verwerpen we de nulhypothese van normaliteit op significantieniveau $\alpha = 0.05$. Is de p-waarde groter dan 0.05, kunnen we de normaliteit niet verwerpen op significantieniveau $\alpha = 0.05$.

Vooraleer de test uit te voeren, kan je best een aantal grafische voorstellingen van de gegevens maken: boxplot, histogram, QQ-plot.

De Shapiro-Wilk test voeren we in R met de volgende code uit:

```
shapiro.test(datajuni$TEMPERATUUR)
```

Opmerking: Extra informatie over de Shapiro-Wilk test (gebruik en syntax) in R kan je vinden door `help(shapiro.test)` in te geven.

Oefening 4 Onderzoek de normaliteit van de variabele temperatuur. Staaft deze formele Shapiro-Wilk test je besluit omtrent normaliteit uit Oefening 1?

.....

.....

Merk op: In de praktijk ga je eerst de normaliteit van de te onderzoeken variabele na om dan pas een hypothesetest uit te voeren. Als de variabele niet normaal verdeeld is, heeft het resultaat van de hypothesetest immers geen enkele betekenis!

3 Test over de mediaan

Gegeven een steekproef $X_1, \ldots, X_n$ met continue verdeling F. Dan kunnen we een hypothesetest uitvoeren omtrent de mediaan van de onbekende verdeling F als volgt:

$$\begin{cases} H_0 : med(F) = \mu_0 \\ H_1 : med(F) \neq \mu_0 \end{cases}$$

Als teststatistiek neemt men $T = \sum_i^n Z_i$ met $Z_i = 1$ als $X_i > \mu_0$. Dan is $T \sim_{H_0} Bin(n, 0.5)$.

In R zal je eerst het BSDA pakket moeten installeren. Dit doe je via Packages/Install packages. Elke keer als je R opent, moet je het pakket BSDA uploaden (bv. via Packages/Load package). De mediaantest (tekentest) voer je dan uit met het commando
SIGN.test(x,md=μ_0). **Opmerking:** Voor meer informatie typ **?SIGN.test** in R.

Voorbeeld: Men wil nagaan of een bepaald middel het gewicht doet afnemen. Bij 20 proefpersonen vond men dat het gewicht bij 13 personen na het volgen van de kuur kleiner was geworden. Stel X de gewichtsname. Het is wenselijk dat X positief is.

De vector met gewichtsverliezen wordt gegeven door

$$x = c(c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3),$$
$$c(-0.1, -0.2, -0.3, -0.4, -0.5, -0.6, -0.7)).$$

We toetsen $H_0 : med(X) = 0$ versus $H_1 : med(X) > 0$.
In R typen we vervolgens **SIGN.test(x, md = 0, alternative = "greater")**.

Oefening 4 Gebruik opnieuw de dataset *datajuni.xls* over temperaturen in juni. Op het weerbericht beweert men dat de mediane junitemperatuur op het middaguur in Leuven 20 graden Celsius is. Toets deze bewering.

- Wat is de nul- en alternatieve hypothese?
  .....
  .....
- Wat is de teststatistiek en zijn verdeling onder H_0?
  .....
  .....
- Schrijf je R-code uit.
  .....
  .....
  .....
  .....
  .....
- Wat besluit je?
  .....
  .....

4 Test over percentage p

Gegeven een steekproef $X_1, \ldots, X_n \sim Bin(1, p)$. Stel dat we volgende hypothese willen testen over de succeskans p met $0 < p_0 < 1$:

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases}$$

Als teststatistiek neemt men $T = \sum_i^n X_i$. Dan is $T \sim_{H_0} Bin(n, p_0)$ en hieruit kan men eenvoudig het aanvaardingsgebied en de p-waarde bepalen.

In R gebruiken we de functie **binom.test(x,n,p)**. **Opmerking:** Voor meer informatie typ **?binom.test** in R.

Oefening 5 Beschouw een populatie van 30 personen. Men stelt $X = 1$ voor een roker en $X = 0$ voor een niet-roker. Men neemt waar

$$1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1.$$

Toets of er een betekenisvolle voorkeur is voor roken.

- Wat is de nul- en alternatieve hypothese?
  .....
  .....
- Wat is de teststatistiek en zijn verdeling onder H_0?
  .....
  .....

- Schrijf je R-code uit.
  .....
  .....
  .....
  .....
  .....
- Wat besluit je?
  .....
  .....

5 Vergelijken van 2 groepen

5.1 Verschil in gemiddelden van twee normale populaties

Stel dat men beschikt over twee steekproeven $(X_1, X_2, \ldots, X_n)$ en $(Y_1, Y_2, \ldots, Y_m)$ uit twee onafhankelijke normale populaties met $X_i \sim N(\mu_1, \sigma_1^2)$ en $Y_i \sim N(\mu_2, \sigma_2^2)$. We willen een test uitvoeren omtrent het verschil van het gemiddelde van de twee populaties, nl. een test voor $\mu_1 - \mu_2$. De tweezijdige test die we beschouwen is

$$H_0 : \mu_1 - \mu_2 = 0 \text{ versus } H_1 : \mu_1 - \mu_2 \neq 0.$$

We kunnen ook éénzijdig testen bekijken, nl.

$$\begin{array}{lll} H_0 : \mu_1 - \mu_2 = 0 & \text{versus} & H_1 : \mu_1 - \mu_2 > 0 \quad \text{(rechtséénzijdig)} \\ & \text{of} & \\ H_0 : \mu_1 - \mu_2 = 0 & \text{versus} & H_1 : \mu_1 - \mu_2 < 0 \quad \text{(linkséénzijdig)} \end{array}$$

Hier moeten we een onderscheid maken tussen gepaarde en ongepaarde gegevens. Bij gepaarde (niet-onafhankelijke!) gegevens is elke steekproefwaarde x_i verbonden met juist 1 steekproefwaarde y_i. Uiteraard is $n_1 = n_2$ bij gepaarde gegevens. Als X bv. de lengte is van de linkerarm en Y de lengte van de rechterarm gemeten bij $n = 12$ personen, dan beschikken we over gepaarde gegevens. Is X de lengte van de linkerarm bij meisjes en Y de lengte van de linkerarm bij jongens, dan beschikken we over ongepaarde gegevens.

5.1.1 Gepaarde gegevens

Bij gepaarde gegevens ga je over op de toevalsvariabele $V = X - Y$, en dus ook $V_i = X_i - Y_i$. Dan is $\mu_V = \mu_1 - \mu_2$. H_0 wordt nu dus $\mu_V = 0$. Analoog vervangen we bij H_1 ook $\mu_1 - \mu_2$ door μ_V. Als X en Y beide normaal verdeeld zijn, zal V ook een normale verdeling hebben, en wordt dit probleem eigenlijk gereduceerd tot een toets voor één normale variabele. De teststatistiek wordt dus:

$$T = \frac{\bar{V} - \mu_V}{S_V/\sqrt{n}},$$

waarbij $S_V^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(V_i - \bar{V})^2$. Onder de nulhypothese (d.w.z. $\mu_V = 0$) heeft deze teststatistiek een t_{n-1} verdeling.

Het betrouwbaarheidsinterval dat equivalent is met de tweezijdige hypothese wordt gegeven door

$$[\bar{v} - t_{n-1,\alpha/2}\frac{s_V}{\sqrt{n}}, \bar{v} + t_{n-1,\alpha/2}\frac{s_V}{\sqrt{n}}]$$

met $\bar{v} = \bar{x} - \bar{y}$. Wanneer dit betrouwbaarheidsinterval 0 bevat, besluiten we dat μ_1 en μ_2 niet significant verschillen. Dit is equivalent met het niet verwerpen van H_0. Als 0 buiten het betrouwbaarheidsinterval valt, is μ_1 significant verschillend van μ_2.

De p-waarde wordt voor de verschillende alternatieve hypothesen gegeven door:

$$\begin{aligned} p\text{-waarde} &= 2P(T > |\frac{\bar{v}}{s_V/\sqrt{n}}|) \quad \text{als } \mathrm{H}_1 : \mu_1 - \mu_2 = \mu_V \neq 0, \\ &= P(T > \frac{\bar{v}}{s_V/\sqrt{n}}) \quad \text{als } \mathrm{H}_1 : \mu_1 - \mu_2 = \mu_V > 0 \quad \text{en} \\ &= P(T < \frac{\bar{v}}{s_V/\sqrt{n}}) \quad \text{als } \mathrm{H}_1 : \mu_1 - \mu_2 = \mu_V < 0 \end{aligned}$$

met $T \sim t_{n-1}$.

5.1.2 Ongepaarde gegevens waarbij populaties eenzelfde variantie $\sigma_1^2 = \sigma_2^2$ hebben

De teststatistiek die wordt gebruikt als de ongekende populatievarianties dezelfde zijn, is

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2},$$

met S_p de vierkantswortel uit de gepoolde variantie: $S_p^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$.
Onder H_0 is $\mu_1 = \mu_2$ zodat:

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim_{H_0} t_{n_1+n_2-2}.$$

Het betrouwbaarheidsinterval dat equivalent is met de tweezijdige hypothese wordt gegeven door

$$[\bar{x} - \bar{y} - t_{n_1+n_2-2,\alpha/2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{x} - \bar{y} + t_{n_1+n_2-2,\alpha/2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}]$$

Wanneer dit betrouwbaarheidsinterval 0 bevat, besluiten we dat μ_1 en μ_2 niet significant verschillen. Dit is equivalent met het niet verwerpen van H_0. Als 0 buiten het betrouwbaarheidsinterval valt, is μ_1 significant verschillend van μ_2.

De p-waarde wordt voor de verschillende alternatieve hypothesen gegeven door:

$$\begin{aligned} p\text{-waarde} &= 2P(T > |\frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}|) \quad \text{als } \mathrm{H}_1 : \mu_1 - \mu_2 \neq 0, \\ &= P(T > \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}) \quad \text{als } \mathrm{H}_1 : \mu_1 - \mu_2 > 0 \quad \text{en} \\ &= P(T < \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}) \quad \text{als } \mathrm{H}_1 : \mu_1 - \mu_2 < 0 \end{aligned}$$

met $T \sim t_{n_1+n_2-2}$.

5.1.3 Ongepaarde gegevens waarbij beide populaties een verschillende variantie hebben

Indien de varianties van de twee normale populaties verschillen, wordt een andere teststatistiek gebruikt, namelijk

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

Onder H_0 heeft T een t-verdeling met aantal vrijheidsgraden gelijk aan r met

$$r = \frac{(s_1^2/n_1 + s_2^2/n_2)}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}.$$

Het hoeft dus niet te verbazen dat het aantal vrijheidsgraden geen natuurlijk getal is.

5.1.4 Een test om na te gaan of de varianties gelijk zijn

Volgende test kan worden toegepast in R zodat je kan besluiten welke van de twee testen je moet gebruiken om je hypothese na te gaan:

$$\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2 \text{ versus } \mathrm{H}_1 : \sigma_1^2 \neq \sigma_2^2.$$

Als teststatistiek gebruikt men S_1^2/S_2^2 die een F_{n_1-1,n_2-1} verdeling bezit.

Bij deze test wordt een p-waarde gegeven in R .

- Indien deze p-waarde $< \alpha$ zullen we de gelijkheid van varianties verwerpen en dus de test met verschillende varianties uitvoeren.
- Als de p-waarde $> \alpha$ zullen we de gelijkheid van varianties niet verwerpen en veronderstellen dat de varianties van beide variabelen gelijk zijn. We zullen dan ook de test met gelijke varianties uitvoeren.

Hoe al deze testen precies werken in R wordt uitgelegd m.b.v. volgende oefeningen.

5.2 Cholesterolgehalte

Voor deze oefening gebruik je de dataset *cholesterol.txt*. In een Amerikaans medisch centrum is er een studie gehouden naar het cholesterolgehalte in het bloed van hartaanvalpatiënten. Het cholesterolgehalte van 28 hartaanvalpatiënten werd eerst na 2 dagen gemeten (*Cholest2*) en vervolgens nog eens na 4 dagen (*Cholest4*). Ga na of het gemiddelde cholesterolgehalte in het bloed 2 dagen na een hartaanval significant verschillend is van het cholesterolgehalte in het bloed 4 dagen na een hartaanval.

Oefening 7

1. Beschik je over gepaarde of ongepaarde gegevens?

   .....
   .....

2. Formuleer de hypothese die je wilt testen (d.w.z. nulhypothese en alternatieve hypothese!).

   .....
   .....

3. Wat zijn de onderstellingen van de bijhorende test?

   .....
   .....

4. Welke teststatistiek ga je gebruiken?

   .....
   .....

   Dit soort test kunnen we in R als volgt oplossen:

   ```
   cholest2=cholesterol$cholest2
   cholest4=cholesterol$cholest4
   t.test(cholest2,cholest4,paired=TRUE)
   ```

5. Wat is de experimentele waarde?

   .....
   .....

6. Hoe is de p-waarde gedefinieerd en waaraan is de p-waarde gelijk? Maak ook een schets van de p-waarde.

   .....

   .....
   .....
   .....
   .....

7. Wat besluit je nu op significantieniveau $\alpha = 0.05$?

   .....
   .....

Oefening 8 Indien we willen testen of het gemiddelde cholesterolgehalte in het bloed 2 dagen na een hartaanval significant **groter** is dan het cholesterolgehalte in het bloed 4 dagen na een hartaanval, wat besluit je dan?

1. Formuleer de hypothese die je wilt testen.

   .....
   .....

2. Hoe is hier de p-waarde gedefinieerd? Maak ook een schets.
   .....

   .....
   .....
   .....
   .....

3. Wat besluit je?

   .....

   .....

5.3 Evenwicht

Negen oudere en acht jonge mensen werd gevraagd deel te nemen aan een evenwichtsexperiment. Elke persoon stond op een platform en moest na het horen van een zeker geluid dat op een willekeurig moment te horen was, zo snel mogelijk op een knop drukken die ze in hun hand vast hielden. De dataset *evenwicht.xls* bevat de gegevens horende bij dit experiment. De variabelen zijn de volgende:

$$\textit{Leeftijd} = \begin{cases} 1 & \text{als de persoon een oudere persoon is,} \\ 2 & \text{als de persoon jong is.} \end{cases}$$

De tweede variabele is *zwaai*, nl. het logaritme van de grootte van de zwaai in millimeters die de persoon naar voren en achteren maakt bij het drukken op de knop.

We willen de hypothese testen dat leeftijd geen invloed heeft op de gemiddelde logaritmische zwaai van een persoon. Stel μ_1 is de gemiddelde logaritmische zwaai van een oude persoon en μ_2 is de gemiddelde logaritmische zwaai van een jonge persoon. De hypothese die we willen testen is dan

$$\mathrm{H}_0 : \mu_1 = \mu_2 \text{ versus } \mathrm{H}_1 : \mu_1 \neq \mu_2.$$

Oefening 9

1. Hebben we gepaarde of ongepaarde gegevens?

   .....
   .....

Als de variabele *zwaai* normaal verdeeld is voor de twee groepen, kunnen we de t-test uitvoeren. We moeten echter eerst nagaan of de varianties voor beide groepen gelijk zijn. De R -code hiervoor is:

```
zwaai=evenwicht$zwaai
leeftijd=evenwicht$leeftijd
var.test(zwaai[leeftijd==1],zwaai[leeftijd==2])
```

De code voor de t-test is dan

```
t.test(zwaai[leeftijd==1],zwaai[leeftijd==2],var.equal=TRUE)
```

als we niet verworpen hebben dat de varianties gelijk zijn, in het andere geval gebruik je `var.equal=FALSE`.

De resultaten van de test $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_1 : \sigma_1^2 \neq \sigma_2^2$ vinden we terug in Tabel 1. De p-waarde is gelijk aan 0.3779.

F	num df	denom df	p-value
1.9961	8	7	0.3779

Tabel 1: Resultaten bij de test of $\sigma_1^2 = \sigma_2^2$ bij dataset *evenwicht*.

De resultaten van de t-test staan samengevat in Tabel 2.

t	df	p-value
2.6315	15	0.01888

Tabel 2: Resultaten bij de t-test bij dataset *evenwicht*.

Aangezien de hypothese dat de varianties gelijk zijn niet wordt verworpen, wordt de experimentele waarde of t-waarde gegeven door 2.6315 en is de test uitgevoerd met $15 = n_1 + n_2 - 2 = 9 + 8 - 2$ vrijheidsgraden.

Opmerking: Extra informatie over de F-test (test voor gelijkheid van de varianties) in R kan je vinden door `help(var.test)` in te geven.

Oefening 10

1. Welke teststatistiek gebruik je en welke verdeling heeft deze teststatistiek onder H_0?

   .....
   .....

2. Hoe is de p-waarde gedefinieerd?

   .....
   .....

De p-waarde wordt gegeven door $0.01888 < 0.05$. Op significantieniveau $\alpha = 0.05$ wordt de hypothese dat de gemiddelde logaritmische grootte van de zwaai hetzelfde is voor oude en jonge mensen verworpen. Er bestaat dus een verschil in de grootte van de zwaai die mensen maken naargelang hun leeftijd.

Oefening 11 Test dezelfde hypothese op significantieniveau $\alpha = 0.01$. Wat besluit je?

.....
.....
.....
.....
.....

5.4 Lengte van rivieren in Nieuw-Zeeland

De dataset *rivieren.xls* bevat de lengte van 38 rivieren in Nieuw-Zeeland. De lengte wordt weergegeven in kilometer. Dit staat onder de variabele *lengte*. De tweede variabele *zee* is gelijk aan 1 als de rivier stroomt naar de Stille Zuidzee en is gelijk aan 2 als de rivier stroomt naar de Tasmaanse Zee. We gaan na of er een verschil bestaat tussen de gemiddelde lengte van een rivier die naar de Stille Zuidzee stroomt (μ_1) en de gemiddelde lengte van een rivier die naar de Tasmaanse Zee stroomt (μ_2). We testen bijgevolg volgende hypothese:

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_1 : \mu_1 \neq \mu_2.$$

Oefening 12

1. Zijn de gegevens gepaard?

   .....
   .....

2. Ga in de eerste plaats de veronderstelling van de test na.

   .....

   .....

3. Onderzoek de gelijkheid van de varianties op significantieniveau $\alpha = 0.05$.

.....
.....

4. Welke teststatistiek gebruik je en hoe is deze verdeeld onder H_0?

.....
.....

5. Test deze hypothese. Wat besluit je?

.....

.....
.....
.....
.....

6. Test de eenzijdige hypothese $\mathrm{H}_0 : \mu_1 = \mu_2$ vs. $\mathrm{H}_1 : \mu_1 > \mu_2$.

   (a) Hoe is de p-waarde gedefinieerd?

   .....
   .....

   (b) Welke p-waarde bekom je?

   .....
   .....

   (c) Wat besluit je?

   .....
   .....