

H5 - Betrouwbaarheidsintervallen voor een verschil van gemiddelden

Benjamin Huyghe

Statistiek I

1 Inleiding

In dit hoofdstuk bekijken we telkens twee studiestudies met gemiddelden μ_1 en μ_2 , waarbij we verlangen deze gemiddelden te vergelijken met elkaar. In praktijk komen dit soort studies vaak voor. Denk bijvoorbeeld aan een groep patiënten met behandeling en een controlegroep met een placebo.

Om te kunnen bepalen welke toets we kunnen gebruiken om de gemiddelden te vergelijken moeten we weten of de data **gepaard** of **ongepaard** zijn.

Beschouw twee steekproeven: $X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2}$.

Onderstel onafhankelijkheid binnen de groepen: $X_i \perp\!\!\!\perp X_j$ en $Y_i \perp\!\!\!\perp Y_j$ voor $i \neq j$.

- **Ongepaard:** Observaties uit beide groepen onafhankelijk: $X_i \perp\!\!\!\perp Y_i$.
- **Gepaard:** Observaties tussen beide groepen paarsgewijs afhankelijk. We nemen $(X_1, Y_1), \dots, (X_n, Y_n)$ als afhankelijke paren.

Vb. Onderzoekt men de werking van oogdruppels op het linker- en respectievelijk rechteroog van patiënten zal er duidelijk afhankelijkheid optreden dus ogen van dezelfde patiënt. Men spreekt hier dus van gepaarde data.

2 Gepaarde data

Assumpties:

- Gelijke steekproefgroottes: $n := n_1 = n_2$
- $(X_i, Y_i) \perp\!\!\!\perp (X_j, Y_j)$ voor $i \neq j$
- $X_i \sim N(\mu_1, \sigma_1^2)$ i.i.d. en $Y_i \sim N(\mu_2, \sigma_2^2)$ i.i.d.

We kunnen lastig de twee gemiddelden apart bestuderen. We nemen per paar het verschil en kijken naar het gemiddelde verschil. Definieer voor elk paar het verschil $V_i := X_i - Y_i$. Door de onafhankelijkheid tussen de paren zullen dus

ook de verschillen onafhankelijk zijn: $V_i \perp V_j$.
 Hierdoor volgt uit **stelling 3.29** dat

$$\text{Var}(\overline{V_n}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i - Y_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i - Y_i) = \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n}.$$

We kunnen zonder verwarring van $\sigma^2 := \text{Var}(X_i - Y_i)$ spreken omdat alle X_i en Y_i dezelfde verdeling volgen. Deze berekening motiveert volgende aanname. Neem aan dat $V_i \sim N(\mu, \sigma^2)$ met $\mu := \mu_1 - \mu_2$ en $\sigma^2 := \text{Var}(X_i - Y_i)$. Volgens **stelling 4.15** is dan

$$\overline{V_n} \sim N\left(\mu, \frac{\sigma^2}{n}\right). \quad (1)$$

Merk op dat we deze normaliteitsaanname niet nodig hebben indien steekproefgrootte n voldoende groot (bijvoorbeeld $n \geq 30$) is, in dit geval geldt (1) ook door de **centrale limietstelling** (zie hoofdstuk 7).

We hebben dus onze onderzoeksvraag beperkt tot één steekproef V_i , waarvan $\overline{V_n}$ een schatter voor μ is. We kunnen met behulp van de theorie uit **hoofdstuk 4** een betrouwbaarheidsinterval voor dit gemiddelde opstellen.

2.1 σ gekend

Door (1) is

$$\frac{\overline{V_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1). \quad (2)$$

Via deze verdeling kunnen we eenvoudig een $(1-\alpha)100\%$ -betrouwbaarheidsinterval voor μ opstellen.

$$\begin{aligned} (1-\alpha)100\% &= P\left(-z_{\frac{\alpha}{2}} < \frac{\overline{V_n} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}\right) \\ &= P\left(-\overline{V_n} - \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}} < -\mu < -\overline{V_n} + \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}\right) \\ &= P\left(\overline{V_n} - \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}} < \mu < \overline{V_n} + \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}\right). \end{aligned}$$

We krijgen dus volgend betrouwbaarheidsinterval voor $\mu = \mu_1 - \mu_2$:

$$\left[\overline{X_n} - \overline{Y_n} - \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}, \overline{X_n} - \overline{Y_n} + \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}} \right]$$

2.2 σ niet gekend

Hoewel voorgaande theorie uiteraard nog steeds algemeen geldig is, kent men in praktijk vaak helaas $\sigma = \text{Var}(X_i - Y_i)$ niet. We kunnen wel de steekproef-schatter $S_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - Y_i - (\overline{X_n} - \overline{Y_n}))^2}$ gebruiken.

In **hoofdstuk 4** hebben we namelijk aangetoond dat

$$T = \frac{\overline{V}_n - \mu}{\frac{S_n}{\sqrt{n}}} \sim t_{n-1}. \quad (3)$$

Geheel analoog aan het vorige geval vinden we zo volgend $(1-\alpha)100\%$ -betrouwbaarheidsinterval voor μ :

$$\left[\overline{X}_n - \overline{Y}_n - \frac{S_n}{\sqrt{n}} t_{n-1; \frac{\alpha}{2}}, \overline{X}_n - \overline{Y}_n + \frac{S_n}{\sqrt{n}} t_{n-1; \frac{\alpha}{2}} \right]$$

3 Ongepaarde data

Merk als eerste op dat voor ongepaarde data de steekproefgroottes n_1 en n_2 niet noodzakelijk even groot hoeven te zijn. In praktijk zal de controlegroep in medisch onderzoek vaak kleiner zijn dan de behandelde groep patiënten.

Assumpties:

- $X_i \perp Y_j$ voor alle i, j
- $X_i \sim N(\mu_1, \sigma_1^2)$ i.i.d. en $Y_j \sim N(\mu_2, \sigma_2^2)$ i.i.d.
- Homoscedasticiteit: $\sigma := \sigma_1 = \sigma_2$

De aanname van homoscedasticiteit is nogal riskant en in praktijk zelden voldaan. Om deze voorwaarde na te gaan kan men de boxplots van beide steekproeven bekijken. Als deze ongeveer even groot zijn kan men aannemen dat ongeveer $\sigma_1 \approx \sigma_2$. Voor een betrouwbaardere methode verwijzen we naar Levene's test of de Brown-Forsythe test (zie Statistiek II).

3.1 Schatten van de variantie σ^2

Zoals vaak de realiteit blijkt nemen we hier aan dat we variantie σ^2 niet kennen. We wensen haar dus te schatten via haar steekproefschatter. Aangezien $n_1 \neq n_2$ kunnen we niet op de klassieke definitie van S_n steunen, we zullen een **gepoolde steekproefvariantie** moeten definiëren.

$$S_p'^2 := \frac{\sum_{i=1}^{n_1} (X_i - \overline{X})^2 + \sum_{j=1}^{n_2} (Y_j - \overline{Y})^2}{n_1 + n_2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2} \quad (4)$$

Hoewel deze definitie natuurlijk aanvoelt kunnen we eenvoudig bewijzen dat dit een vertekende schatter voor σ^2 is.

$$\mathbb{E}(S_p'^2) = \frac{1}{n_1 + n_2} ((n_1 - 1) \mathbb{E}(S_1^2) + (n_2 - 1) \mathbb{E}(S_2^2)) = \frac{n_1 + n_2 - 2}{n_1 + n_2} \sigma \neq \sigma$$

Om te garanderen dat deze schatter wél onvertekend is introduceren we volgende definitie.

Definitie 1 (Gepoolde steekproefvariantie). De *gepoolde steekproefvariantie* voor steekproeven $X_1, \dots, X_{n_1}$ en $Y_1, \dots, Y_{n_2}$ wordt gegeven door

$$S_p^2 := \frac{\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{j=1}^{n_2} (Y_j - \bar{Y})^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}. \quad (5)$$

3.2 Opstellen van het betrouwbaarheidsinterval

3.2.1 Verdeling van de gepoolde steekproefvariantie

Om een betrouwbaarheidsinterval voor $\mu := \mu_1 - \mu_2$ te vinden moeten we de verdeling van S_p^2 kennen.

Uit **stelling 4.23** weten we dat $(n_1 - 1)S_1^2 \sim \sigma_1^2 \chi_{n_1-1}^2$ en $(n_2 - 1)S_2^2 \sim \sigma_2^2 \chi_{n_2-1}^2$.

Lemma 1.

$$\chi_{n_1-1}^2 + \chi_{n_2-1}^2 = \chi_{n_1+n_2-2}^2$$

Bewijs. Volgens de definitie van een chi-kwadraat-verdeling is $\chi_n^2 = \sum_{i=1}^n Z_i^2$ met $Z_i \sim N(0, 1)$ i.i.d. Hierdoor is dus ook

$$\chi_{n_1-1}^2 + \chi_{n_2-1}^2 = \sum_{i=1}^{n_1-1} Z_i^2 + \sum_{j=1}^{n_2-1} Z_j^2 = \sum_{i=1}^{n_1+n_2-2} Z_i^2 = \chi_{n_1+n_2-2}^2.$$

□

Door dit lemma vinden we met andere woorden dat

$$S_p^2 \sim \frac{\sigma_1^2 \chi_{n_1-1}^2 + \sigma_2^2 \chi_{n_2-1}^2}{n_1 + n_2 - 2} = \sigma^2 \frac{\chi_{n_1-1}^2 + \chi_{n_2-1}^2}{n_1 + n_2 - 2} = \frac{\sigma^2 \chi_{n_1+n_2-2}^2}{n_1 + n_2 - 2}.$$

Of nog

$$\frac{S_p^2}{\sigma^2} \sim \frac{\chi_{n_1+n_2-2}^2}{n_1 + n_2 - 2}. \quad (6)$$

3.2.2 Verdeling van het verschil van de steekproefgemiddelden

Aangezien $n_1 \neq n_2$ kunnen we niet voor elk paar een verschil V definiëren. We kunnen echter wel nog $\mu := \mu_1 - \mu_2$ afschatten met behulp van haar steekproef-schatter $\bar{V} := \bar{X} - \bar{Y}$.

Opnieuw kunnen we door de onafhankelijkheid bij ongepaarde data haar variantie vinden als

$$\begin{aligned} \text{Var}(\bar{V}) &= \text{Var}\left(\frac{1}{n_1} \sum_{i=1}^{n_1} X_i - \frac{1}{n_2} \sum_{j=1}^{n_2} Y_j\right) = \text{Var}\left(\frac{1}{n_1} \sum_{i=1}^{n_1} X_i\right) + \text{Var}\left(-\frac{1}{n_2} \sum_{j=1}^{n_2} Y_j\right) \\ &= \frac{1}{n_1^2} \sum_{i=1}^{n_1} \text{Var}(X_i) + \frac{1}{n_2^2} \sum_{j=1}^{n_2} \text{Var}(Y_j) = \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sigma_1^2 + \frac{1}{n_2^2} \sum_{j=1}^{n_2} \sigma_2^2. \end{aligned}$$

Waardoor dus

$$\text{Var}(\bar{V}) = \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \sigma^2. \quad (7)$$

Door aan te nemen dat $\bar{V} \sim N\left(\mu, \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \sigma^2\right)$ of door de centrale limietstelling (voor n groot) vinden we dan dat

$$\frac{\bar{V} - \mu}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1). \quad (8)$$

3.2.3 Opstellen van het betrouwbaarheidsinterval

Als we dan de resultaten uit (6) en (8) samenleggen vinden we dat

$$T = \frac{\bar{V} - \mu}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\frac{\bar{V} - \mu}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}{\sqrt{\frac{S_p^2}{\sigma^2}}} \sim \frac{N(0, 1)}{\sqrt{\frac{\chi_{n_1+n_2-2}^2}{n_1+n_2-2}}} = t_{n_1+n_2-2}. \quad (9)$$

Ook bij ongepaarde data kunnen we dus nu een $(1-\alpha)100\%$ -betrouwbaarheidsinterval vanuit deze student-t-verdeling opstellen.

$$\begin{aligned} (1-\alpha)100\% &= P\left(-t_{n_1+n_2-2; \frac{\alpha}{2}} < \frac{\bar{V} - \mu}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{n_1+n_2-2; \frac{\alpha}{2}}\right) \\ &= P\left(\bar{V} - S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{n_1+n_2-2; \frac{\alpha}{2}} < \mu < \bar{V} + S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{n_1+n_2-2; \frac{\alpha}{2}}\right). \end{aligned}$$

We verkrijgen dus een $(1-\alpha)100\%$ -betrouwbaarheidsinterval voor $\mu := \mu_1 - \mu_2$:

$$\left[\bar{X} - \bar{Y} - S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{n_1+n_2-2; \frac{\alpha}{2}}, \bar{X} - \bar{Y} + S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{n_1+n_2-2; \frac{\alpha}{2}} \right]$$