Verklarende statistiek in SPSS

Standaard: Tweezijdige test → Tweezijdige p-waarde (= Kans dat t-verdeelde variabele minstens even ver verwijderd is van 0 als waargenomen teststatistiek als H_0 waar is)

Vergelijken van gemiddelden voor twee groepen

Ongepaarde t-test

- Voor gelijke varianties van groepen
- Voor ongelijke varianties van groepen

Keuze maken: *Levene's Test for Equality of Variances* → H_0 = Varianties zijn gelijk

- *p-waarde < 5 %* → H_0 verwerpen → Varianties zijn niet gelijk
- *p-waarde > 5 %* → H_0 niet verwerpen → Varianties verschillen niet sterk genoeg

VW'en:

- *Elke groep is normaal verdeeld* → Elk groepgemiddelde is normaal verdeeld
  Niet voldaan aan VW: Steekproefgemiddelde van elke groep is bij benadering normaal verdeeld als steekproefgrootte ≥ 30 is
  → **Nadeel**: Testen maken meer type I fouten (te snel H_0 verwerpen)
  Voorzichtig zijn met conclusies als:
  - ❖ Steekproef niet heel groot is
  - ❖ Er afwijkingen van normaliteit zijn
  - ❖ Er zwak bewijs tegen H_0 is (p-waarde tussen 1 en 5 %)
- *Groepen zijn onafhankelijk van elkaar* = Moeilijk te testen

Normaliteit nagaan: *Analyze* → *Descripitive Statistics* → *Explore* → Kies *Dependent List* en *Factor List* → Kies bij *Display* optie *Plots* → Vink bij *Plots* optie *Normality plots with tests* aan → Vink bij *Plots* andere opties uit of selecteer *None* → *Continue* → *OK*

- *QQ-plot* = Geobserveerde kwantielen tegenover theoretische kwantielen onder normaliteit
  - ❖ Punten liggen op lijn → Normale verdeling
  - ❖ Punten liggen niet op lijn → Geen normale verdeling
- *Shapiro-Wilk test* → H_0 = Normale verdeling
  - ❖ Teststatistiek is te extreem → Afwijking van normaliteit is te sterk

  → Meer power als steekproefgrootte stijgt

t-test uitvoeren: *Analyze* → *Compare Means* → *Independent Samples T Test* → Kies *Test Variable* en *Grouping Variable* → *Define Groups* → Geef waarden in voor elke groep → *Continue* → *OK*

Output:

- Group Statistics
- Independent Samples Test

Eenzijdig testen

Manueel eenzijdige p-waarde afleiden uit output SPSS

p-waarde = Kans dat we minstens even extreme teststatistiek uitkomen als degene die we waaarnemen als H_0 waar is

T = Teststatistiek

t_k = t-verdeling die T volgt als H_0 waar is → Symmetrisch rond 0

t = Waarde teststatistiek voor steekproef

Tweezijdige p-waarde:

$$p = P(T < -|t| \cup |t| < T) = P(T < -|t|) + P(|t| < T) = 2P(T < -|t|) = 2P(|t| < T)$$

Twee groepen A en B → Teken van t = Teken van $\bar{x}_A - \bar{x}_B$:

- **$H_0: \mu_A = \mu_B$**
  $H_1: \mu_A > \mu_B$
  Eenzijdige p-waarde = Kans dat T minstens even groot is als t als H_0 waar is = $P(T \geq t)$
  - $\bar{x}_A > \bar{x}_B \rightarrow t > 0 \rightarrow$ Steekproef ondersteunt H_1
    Tweezijdige p-waarde: $p = 2P(|t| < T) = 2P(t < T)$
    → Eenzijdige p-waarde: $P(t < T) = \frac{p}{2}$
  - $\bar{x}_A < \bar{x}_B \rightarrow t < 0 \rightarrow$ Steekproef ondersteunt H_0
    Tweezijdige p-waarde: $p = 2P(T < -|t|) = 2P(T < t) = 2(1 - P(t < T))$
    → Eenzijdige p-waarde: $P(t < T) = 1 - \frac{p}{2}$
- **$H_0: \mu_A = \mu_B$**
  $H_1: \mu_A < \mu_B$
  Eenzijdige p-waarde = Kans dat T maximaal gelijk is aan t als H_0 waar is = $P(T < t)$
  - $\bar{x}_A > \bar{x}_B \rightarrow t > 0 \rightarrow$ Steekproef ondersteunt H_0
    Tweezijdige p-waarde: $p = 2P(|t| < T) = 2P(t < T) = 2(1 - P(T < t))$
    → Eenzijdige p-waarde: $P(T < t) = 1 - \frac{p}{2}$
  - $\bar{x}_A < \bar{x}_B \rightarrow t < 0 \rightarrow$ Steekproef ondersteunt H_1
    Tweezijdige p-waarde: $p = 2P(T < -|t|) = 2P(T < t)$
    → Eenzijdige p-waarde: $P(T < t) = \frac{p}{2}$

Samenvatting:

- *Steekproef ondersteunt H_1* → $\text{éénzijdige p} - \text{waarde} = \frac{tweezijdige\ p-waarde}{2}$
- *Steekproef ondersteunt H_1 niet* → $\text{éénzijdige p} - \text{waarde} = 1 - \frac{tweezijdige\ p-waarde}{2}$

Gepaarde t-test

= t-test voor één groep
VW: Verschillen van gegevens zijn normaal verdeeld of er zijn voldoende observaties zodat gemiddelde normaal verdeeld is
Nadeel: Aantal type I fouten neemt toe als steekproef niet heel groot is of als er inbreuken zijn op normaliteit → Voorzichtige conclusies trekken als p-waarde tussen 1 en 5 % ligt
Normaliteit testen:

- Variabele die verschil tussen gepaarde observaties voorstelt, aanmaken: *Transform → Compute Variable* → Geef naam bij *Target Variable* → Geef formule bij *Numeric Expression* → *OK*
- Normaliteit testen: *Analyze → Descriptive statistics → Explore* → Kies *Dependent List* → Kies bij *Display* optie *Plots* → Vink bij *Plots* optie *Normality plots with tests* aan → Vink bij *Plots* andere opties uit of selecteer *None* → *Continue* → *OK*
- Gepaarde t-test uitvoeren:
  - *Analyze → Compare Means → Paired-Samples T-test* → Selecteer *Variable1* en *Variable2* → *OK*
  - *Analyze → Compare Means → One-Sample T Test → Kies Test Variable → OK*

Anova

= Gemiddelden van meerdere groepen tegelijk vergelijken
= Alternatief voor tussen elke twee groepen paarsgewijs ongepaarde t-test uitvoeren (kans om in al die testen minstens één type I fout te maken > 5 %)
= Vrij robuust tegen inbreuken op gelijkheid van varianties als groepgroottes ± gelijk zijn → Voorzichtige conclusies trekken als p-waarde tussen 1 en 5 % ligt
H_0: Alle groepgemiddelden zijn gelijk aan elkaar
H_1: Minstens twee groepen hebben ander gemiddelde

H_0 verwerpen → Post hoc analyse (paarsgewijze vergelijking via t-testen) om te onderzoeken welke groepen verschillen van gemiddelde
Correctie op p-waarden (Tukey) om op type I fouten te controleren
VW'en:

- Normaliteit van gegevens in elke groep
- Gelijkheid van varianties van elke groep
- Onafhankelijkheid van groepen onderling

Nadeel: Aantal type I fouten neemt toe als steekproef niet heel groot is of als er inbreuken zijn op normaliteit → Voorzichtige conclusies trekken als p-waarde tussen 1 en 5 % ligt
Normaliteit testen: *Analyze* → *Descriptive statistics* → *Explore* → Kies *Dependent List* en *Factor List* → Kies bij *Display* optie *Plots* → Vink bij *Plots* optie *Normality plots with tests* aan → Vink bij *Plots* andere opties uit of selecteer *None* → *Continue* → *OK*
Nagaan of varianties gelijk zijn via Levene test, ANOVA-test uitvoeren en post hoc analyse uitvoeren: *Analyze* → *Compare Means* → *One-Way ANOVA* → Kies *Dependent List* en *Factor* → *Options* → Vink *Homogeneity of variance test* aan → *Continue* → *Post Hoc* → Vink *Tukey* aan → *Continue* → *OK*
Output:

- Test of Homogeneity of Variances
- ANOVA
- Post Hoc Tests
- Homogeneous Subsets

Probleem: Tekstvariabelen staan niet in de lijst bij *One-Way ANOVA*
→ **Oplossing**: Numerieke variabele aanmaken uit tekstvariabele: *Transform* → *Automatic Recode* → Kies variabele en geef nieuwe naam → *Add New Name* → *OK*
Exacte grootte van groepen berekenen: *Analyze* → *Descriptive Statistics* → *Frequencies*

Multilineaire regressie

De theorie

Lineaire model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon$ met $\varepsilon \sim N(0,\sigma^2)$
Assumpties:

- *Lineariteit*: $Y_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_n X_{n,i} + \varepsilon_i$
  met $E(\varepsilon_i | X_{1,i}, \dots, X_{n,i}) = E(Y_i - (\beta_0 + \beta_1 X_{1,i} + \cdots + \beta_n X_{n,i}) | X_{1,i}, \dots, X_{n,i}) = 0$
  → Onvertekende schatters $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_n$ voor $\beta_0, \beta_1, \dots, \beta_n$ via kleinste kwadratenmethode
- *Homoscedasticiteit*: $Var(\varepsilon_i | X_{1,i}, \dots, X_{n,i}) = \sigma^2$
- *Normaliteit*: $\varepsilon_i \sim N(\bullet,\bullet)$
- *Onafhankelijke storingstermen*: $\forall\, i, j \; met \; i \neq j: Cov(\varepsilon_i, \varepsilon_j | X_{1,i}, X_{1,j}, \dots, X_{n,i}, X_{n,j}) = 0$

Onvertekende schatter $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \cdots \hat{\beta}_n X_{n,i}$ voor responsvariabele Y_i doordat $E(Y_i | X_1, \dots, X_n) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$
Eigenschap: $E(Y | X_1, \dots, X_k + 1, \dots, X_n) - E(Y | X_1, \dots, X_k, \dots, X_n) = \beta_k$
→ **Interpretie van coëfficiënt β_k**:

- X_k = Schaalvariabele: Verhoging van X_k met één eenheid gaat gemiddeld gepaard met verhoging van Y met β_k eenheden
- X_k = Indicatorvariabele: Verschil voor Y tussen individuen waarvoor $X_k = 1$ en individuen waarvoor $X_k = 0$ is gemiddeld gelijk aan β_k

Eigenschap: $E(Y | X_1 = 0, \dots, X_n = 0) = \beta_0$
Schatter $\hat{\beta}_0$ kan sterk afwijken van β_0 als er weinig tot geen individuen zijn met verklarende variabelen in de buurt van $X_1 = 0, \dots, X_n = 0$

De praktijk

Assumpties verifiëren aan de hand van:

- *Residuplot* = Scatterplot = Gestandaardiseerde storingstermen of residu's $\frac{\hat{\varepsilon}_i}{\hat{\sigma}}$ tegenover geschatte of voorspelde waarden $\hat{Y}_i$
  - ❖ **Lineariteitsassumptie geldt** → Puntenwolk is gecentreerd rond x-as
  - ❖ Trend in puntenwolk die afwijkt van x-as → **Lineariteitsassumptie geldt niet**
  - ❖ **Homoscedasticiteitsassumptie geldt** → Verticale spreiding punten rond x-as is overal ± gelijk
  - ❖ Sterkte variatie in spreiding → **Homoscedasticiteitsassumptie geldt niet** = Heteroscedasticiteit
- *PP-plot* = Probability-Probability-plot = Geobserveerde kwantielen tegenover empirische cumulatieve verdelingsfunctie
  → **Voordeel t.o.v. QQ-plot**: PP-plot neemt afwijking in midden gemakkelijker waar
  → **Nadeel t.o.v. QQ-plot**: QQ-plot neemt afwijkingen in staarten gemakkelijker waar
  - ❖ **Normaliteitsassumptie geldt** → Punten liggen op rechte lijn
  - ❖ Punten liggen niet op rechte lijn → **Normaliteitsassumptie geldt niet**

Goed bruikbaar model vinden: Achterwaartse regressie = Starten van volledige model (met alle mogelijke predictoren of verklarende variabelen) en iteratief minst significante predictor (= grootste p-waarde) verwijderen uit model en lineaire model zonder predictor schatten → Herhalen tot alle predictoren significant zijn

Coëfficiënten van model schatten: *Analyze* → *Regression* → *Linear* → Geef afhankelijke variabele in *Dependent* en onafhankelijke variabelen in *Independent* → *Method* → *Enter* → *Plots* → Geef *ZRESID* in *Y* en *ZPRED* in *X* → Vink *Normal Probability Plot* en *Histogram* aan → *Continue* → *OK*

Output:

- *ANOVA*: Om te oordelen of model nuttig is
  $H_0 : \beta_1 = \ldots = \beta_n = 0$
  p-waarde < α → H_0 verwerpen → Minstens 1 van verklarende variabelen heeft significante invloed op responsvariabele
- *Coefficients*:
  - ❖ Schatters voor coëfficiënten van model = *Unstandardized Coefficients*
  - ❖ p-waarde voor elke coëfficiënt
- *Charts*: Opgevraagde plots
  - ❖ PP-plot van storingstermen
  - ❖ Residuplot

We houden constante term of intercept automatisch in model, ook al is die niet significant.

Minst significante verklarende variabele verwijderen uit model: *Analyze* → *Regression* → *Linear* → Verwijder minst significante verklarende variabele uit *Independent* → *OK*