

Variabelen → kwantitatief → continu = lengte, geboortegewicht, opbrengst (staafdiagram) → discreet = aantal ... (linkshandigen, zieke planten, etc.)
 → kwalitatief → nominaal = haarkleur, afstuderrichting, provincie (histogram) → ordinaal = hoogst genoten, opleiding, jaarsalaris (in klassen)

Enkelvoudige aselechte steekproef (EAS) = uit de populatie worden volstrekt willekeurig een aantal eenheden genomen.

Gemiddelde: $\bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n} = \frac{1}{n} \cdot \sum y_i$ Gevoelig voor uitbijters.

Mediaan: Rangschik de waarnemingen van klein naar groot. Mediaan is de waarde met 50% v.d. waarnemingen eronder en 50% erboven. = 50%-punt = 50^e percentiel.
 Niet gevoelig voor uitbijters.

Standaardafwijking = standaarddeviatie: s of σ . $\sqrt{\text{variantie}}$ Gevoelig voor uitbijters.
 Variantie = s^2 : $\frac{(y_1 - \bar{y})^2 + \dots + (y_n - \bar{y})^2}{n-1} = \frac{1}{n-1} \cdot \sum (y_i - \bar{y})^2$ met de GR: let op dat hij deelt door $n-1$ (s_x) en niet door n (σ_x).

$Q_1 = 1^e$ kwartiel = 25^e percentiel
 $Q_2 = 2^e$ kwartiel = 50^e percentiel = mediaan
 $Q_3 = 3^e$ kwartiel = 75^e percentiel
 Interkwartiel afstand = $Q_3 - Q_1$ (IKA)

Het p^e percentiel van een groep van n geordende waarnemingen is die waarde waarvoor hoogstens $p\%$ van de waarnemingen eronder liggen en hoogstens $(100-p)\%$ erboven.

Emperical Rule $(\bar{y}-s, \bar{y}+s)$ bevat ongeveer 68% van de waarnemingen.
 Bij een normale verdeling. $(\bar{y}-2s, \bar{y}+2s)$ bevat ongeveer 95% van de waarnemingen.
 $(\bar{y}-3s, \bar{y}+3s)$ bevat ongeveer 99,7% van de waarnemingen.

Wet van de grote aantallen: Relatieve frequentie stabiliseert zich als men een experiment vele malen herhaalt.

n = Steekproefomvang
 y = Aantal waarnemingen dat aan een bepaalde voorwaarde voldoet.
 p = Kans dat ~~een~~ eenheid aan de voorwaarde voldoet. Schatten d.m.v. y/n .

$A \cup B$ = Uitkomsten die in A voorkomen, in B , of in allebei tegelijk.
 $A \cap B$ = Uitkomsten die zowel in A als in B voorkomen. Als er geen gemeenschappelijke uitkomsten zijn, zijn A en B disjunct. Notatie: $A \cap B = \emptyset$ (lege verzameling).
 \bar{A} = Complement van A = uitkomsten die niet in A voorkomen.

$P(A \cap B) = P(A) \cdot P(B)$ als A en B onafhankelijk zijn.
 $P(A \cup B) = P(A) + P(B)$ als A en B disjunct zijn, want $P(A \cap B)$ is dan ~~0~~ 0.
 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
 $P(\bar{A}) = 1 - P(A)$

Normale verdeling $= y \sim N(\mu, \sigma)$
 Met de GR: ncdf (link.grens, recht.grens, μ, σ)
 Standaard normale verdeling: $N(0, 1)$

Binomiale verdeling $= y \sim B(n, \pi)$
 $P(y=k) = \binom{n}{k} \cdot \pi^k \cdot (1-\pi)^{n-k}$ met de GR: binomcdf(n, π, k)
 Situatie: * n waarnemingen
 * waarnemingen zijn onafhankelijk
 * uitkomst: succes of mislukking
 * kans op succes (π) is voor elke waarneming hetzelfde.

Omrekenen naar st. norm. verdeling: $Z = (y - \mu) / \sigma$
 met y = bovengrens. Dan Z opzoeken in tabel 1 van O&L.
 Daar vind je de linkszijdige kans.
 continue variabelen

discrete variabelen

Discrete toevalsvariabele: $\text{verwachting} = \mu = E(y) = \sum y_i \cdot p_i = \sum \text{uitkomst} \cdot \text{kans}$
 variantie van $y = \sigma^2 = \text{VAR}(y) = \sum (y_i - \mu)^2 \cdot p_i$

Rekenregels voor verwachtingen $\rightarrow Y_{a+by} = a + b \cdot \mu_y$
 $\rightarrow \mu_{x+y} = \mu_x + \mu_y$
voor varianties $\rightarrow \sigma^2_{a+by} = b^2 \cdot \sigma^2_y$
 $\rightarrow \sigma^2_{x+y} = \sigma^2_x + \sigma^2_y$
 $\rightarrow \sigma^2_{x-y} = \text{idem dito}$

Som van trekkingen uit norm. verdeling
 $\Sigma y \sim N(n \cdot \mu_y, \sqrt{n} \cdot \sigma_y)$

Gemiddelde van trekkingen uit norm. ver.
 $\bar{y} \sim N(\mu_y, \frac{\sigma_y}{\sqrt{n}})$ of: $\frac{\text{VAR}(y)}{n}$

Centrale Limietstelling

Wanneer n voldoende groot is ($n \geq 20$) mag je een onbekende verdeling benaderen zoals hierboven staat.

Binomiale toets

$H_0: \pi = \dots$ $H_a: \pi < \text{of} > \text{of} \neq \dots$

P-waarde = De kans dat, berekend onder de aanname dat H_0 waar is, de toetsingsgrootte een waarde zou aannemen die even extreem is als of nog extremer is dan de feitelijk waargenomen uitkomst, in de richting van de H_a .

P-waarde $\leq \alpha \rightarrow H_0$ verwerpen, H_a is aangetoond.

P-waarde $> \alpha \rightarrow H_0$ niet verwerpen, H_a is niet aangetoond.

Bij $H_a: \pi \neq \dots$ doe je een tweezijdige toets. Daarvoor doe je P-waarde $\cdot 2$.

Als H_0 waar is, is er een kans van (ten hoogste) α dat we H_0 toch (ten onrechte) verwerpen en (ten onrechte) concluderen dat H_a juist is.

Normale benadering

$\hat{\pi} \sim N(\pi, \sqrt{\frac{\pi(1-\pi)}{n}})$ met $\hat{\pi} = \frac{y}{n} \rightarrow$ steekproeffractie successen

$y \sim N(n\pi, \sqrt{n\pi(1-\pi)}) \rightarrow$ steekproefaantal successen

Geldig als:
 $n \cdot \pi \geq 5$ én $n \cdot (1-\pi) \geq 5$

Continuïteitscorrectie = Verbetering van de benadering van de discrete verdeling met de continue verdeling.

I.p.v. $P(9 < y < 11)$ te berekenen, neem je $P(8.5 < y < 11.5)$

Toetsen met benadering \rightarrow Toetsingsgrootte: $z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$
 Als H_0 waar is, is $z \sim N(0, 1)$.
 $P(z \geq \text{uitkomst TG}) = \text{P-waarde}$.

! Let op: Boek O&L geeft altijd de linker P-waarde!

Normale verdeling: $y \sim N(\mu, \sigma) \rightarrow$ standaard: $Z \sim N(0, 1)$

Met GR: ncdf(linkergr., rechtergr., μ, σ) $Z = (y - \mu) / \sigma$ "z-score van y"
 =kans Kijk in tabel 1 O&L voor de kans links van z.

Gemiddelde = $\bar{y} \rightarrow \bar{y} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$
 Som = $\sum y \rightarrow \sum y \sim N(n\mu, \sqrt{n}\sigma)$ } Centrale Limietstelling.
 Neem een EAS van omvang n uit een populatie met verwachting μ en standaardafwijking σ . Wanneer n voldoende groot is, zijn de verdelingen bij benadering zo, Onafhankelijk van het type verdeling.

\bar{y} = schatter voor het populatiegemiddelde μ_y
 s = schatter voor de populatie-standaardafwijking $\sigma_y \rightarrow$ op de GR: S_x
 $\sigma_y = \sigma_y = \frac{\sigma_y^2}{\sqrt{n}}$ $se(\bar{y}) = \frac{s}{\sqrt{n}}$ = standaardfout van het gemiddelde.
 (σ = bekend) (σ = onbekend)

Btbh-i $\rightarrow \bar{y} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ of $\bar{y} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$
 tabel 2 O&L voor de z- of t-waarde. Voor σ = bekend kijken bij df. = ∞ (z-verdeling)
 Voor σ = onbekend kijken bij df. = $n-1$ (t-verdeling)

Z-toets

T-toets

- $H_0: \mu = \mu_0$ $H_a: \mu < \text{of } \neq \text{of } > \mu_0$
- TG: $z = \frac{\bar{y} - \mu_0}{\sigma / \sqrt{n}}$
- Onder H_0 is $z \sim N(0, 1)$ verdeeld.
- Onder H_a heeft TG de neiging om kleinere/grotere/grotere of kleinere waarden aan te nemen dan onder H_0 .
- Linkszijdige/rechtszijdige/tweezijdige toets.
- Reken TG uit.
- $P(z \leq \text{uitkomst stap 6})$? Voor tweezijdige toets: neem 2x die waarde.

- Hetzelfde als z-toets.
- TG: $t = \frac{\bar{y} - \mu_0}{s / \sqrt{n}}$
- Onder H_0 heeft t een t-verdeling met $n-1$ df.
- Hetzelfde als z-toets.
- Hetzelfde als z-toets.
- $P(t_{n-1} \geq \text{uitkomst stap 6})$

3) P-waarde $\leq \alpha \rightarrow H_0$ verwerpen. H_a is aangetoond.
 P-waarde $> \alpha \rightarrow H_0$ niet verwerpen. H_a is niet aangetoond.

8) Hetzelfde als z-toets.

Fout van de eerste soort: H_0 verwerpen terwijl H_0 waar is.
 Fout van de tweede soort: H_0 niet verwerpen terwijl H_a waar is.

met GR: tcdf(linkergr., rechtergr., t)
 = P-waarde $n-1$

2 Steekproeven (EAS)

$\sigma_1 = \sigma_2 \rightarrow \bar{y}_1 - \bar{y}_2 \pm t_{\alpha/2} \cdot s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ met $s_p = \frac{n_1 - 1}{n_1 + n_2 - 2} \cdot s_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} \cdot s_2^2$ Btbh-i

\rightarrow uit t-verdeling met df. = $n_1 + n_2 - 2$

- $H_0: \mu_1 - \mu_2 = 0$ $H_a: \mu_1 - \mu_2 < \text{of } \neq \text{of } > 0$
- TG: $t = \frac{\bar{y}_1 - \bar{y}_2 - 0}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
- $t_{n_1 + n_2 - 2}$ -verdeling

~~t-toets~~

- $\sigma_1 \neq \sigma_2 \rightarrow \bar{y}_1 - \bar{y}_2 \pm t_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
 - TG: $t' = \frac{\bar{y}_1 - \bar{y}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
 - t-verdeling met df. volgens SPSS-uitvoer.
- \rightarrow uit t-verdeling met df. volgens SPSS-uitvoer.

Toets van Levene: kijk in SPSS-uitvoer bij Levene's test \rightarrow Sig. $> 0,05$ neem $\sigma_1 = \sigma_2$ (bovenste regel)
 Sig. $\leq 0,05$ neem $\sigma_1 \neq \sigma_2$ (onderste regel)

Btbh-i met SPSS-uitvoer: ondergrens = lower interval of the difference + test value
 bovengrens = upper interval of the difference + test value

Gepaarde t-toets: aan één eenheid uit de steekproef worden meerdere metingen verricht.
 $d = x - y$. t-toets uitvoeren zoals normaal, maar dan met μ_d en σ_d .

① $H_0: \mu_d = D_0 (=0)$

$\rightarrow \bar{d} \pm t_{\alpha/2} \cdot \frac{S_d}{\sqrt{n}}$

② TG: $t = \frac{\bar{d} - D_0}{S_d / \sqrt{n}}$

↳ uit t-verdeling met $df. = n - 1$.

③ t-verdeling met $df. = n - 1$.

Regressie & t-toets voor regressie

$R =$ correlatiecoëfficiënt = Een maat voor de sterkte van de lineaire relatie tussen 2 kwantitatieve variabelen x en y . Correlatie x en $y =$ correlatie y en x .
 Schaal-onafhankelijk. Gevoelig voor uitbijters. $-1 \leq R \leq 1$

$= \frac{1}{n-1} \cdot \sum \left(\frac{x-\bar{x}}{s_x} \right) \left(\frac{y-\bar{y}}{s_y} \right)$

Regressielijn = $\mu_y = \beta_0 + \beta_1 \cdot x$

→ $\beta_1 =$ Verwachte toename in y bij een toename van x van 1 eenheid.

of: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

→ $\beta_0 =$ Verwachte waarde van y bij $x = 0$.

→ Regressie van x op $y \neq y$ op x .

Modelveronderstellingen: 1) ϵ_i is normaal verdeeld als $N(0, \sigma)$ → qq-plot vd. residuen
 2) σ is constant → Residuen diagram → geen patroon, dan klopt het.

$R^2 =$ De door de regressie verklaarde fractie y -variatie.

$s^2 =$ Residuele variantie = Mean Square Residual (SPSS) = $\frac{\sum (\text{Residuen})^2}{n-2}$

Btbi $\beta_0: \hat{\beta}_0 \pm t_{\alpha/2} \cdot se(\hat{\beta}_0)$

→ $\beta_1: \hat{\beta}_1 \pm t_{\alpha/2} \cdot se(\hat{\beta}_1)$

↳ uit t-verdeling met $df. = n - 2$

① $H_0: \beta_i = d$ met $i = 0$ of 1

② TG: $t = \frac{\hat{\beta}_i - d}{se(\hat{\beta}_i)}$

③ t_{n-2} verdeling.

Btbi- i verwachtingswaarde = $\hat{\mu}_y \pm t_{\alpha/2} \cdot se_{\hat{\mu}_y}$ → $df. = n - 2$

Bij $\mu_y = \beta_0 + \beta_1 \cdot x^*$ → puntschatting.

Voorspellingsinterval toekomstige waarneming* = $\bar{y} \pm t_{\alpha/2} \cdot se_{\bar{y}}$ → $df. = n - 2$

Bij $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x^* (= \mu_y)$

$se_{\bar{y}} > se_{\hat{\mu}_y}$ omdat het moeilijker is om een individuele waarneming te voorspellen dan de verwachting.

SPSS-uitvoer

→ PRE = Verwachte y bij een bepaalde x .

→ RES = Residu (ϵ) van een waarneming, voorspelling-waarneming.

→ SEP = Standard Error of Prediction (bij een bepaalde x).

→ LMCI = Lower Mean Confidence Interval } Betrouwbaarheidsinterval*
 UMCI = Upper " " }

→ LICI = Lower Individual Confidence Interval } Voorspellings-*
 UICI = Upper " " } interval

Modelaanname CLS → n moet voldoende groot zijn.

Btbi- i → Variabele is normaal verdeeld in de populatie.

→ Waarnemingen zijn onafhankelijk (EAS)

Z-toets → EAS van omvang n uit een populatie met $y =$ normaal verdeeld met onbekende verwachting μ en bekende σ .

T-toets → Zie z-toets, maar dan met onbekende σ .

2 steekproeven t-toets → EAS van omvang n_1 uit $N(\mu_1, \sigma_1)$ -populatie.

→ " " " n_2 " $N(\mu_2, \sigma_2)$ -populatie

→ eventueel: $\sigma_1 = \sigma_2 = \sigma$

Gepaarde t-toets → Verschillen $d = x - y$ zijn trekkingen uit een $N(\mu_d, \sigma_d)$ -populatie.

Lineaire regressie → Variabele y is voor elke x normaal verdeeld met een verwachting die lineair van x afhangt.

→ Variatie (gemeten door σ) is voor iedere x hetzelfde.

→ Afwijkingen ϵ : zijn onafhankelijk en normaal verdeeld.