statistiek 1

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

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

Gemiddelde: $\bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n} = \frac{1}{n} \cdot \sum y_i$ 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 = 50e percentiel.
Niet gevoelig voor uitbijters.

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

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

Het pe 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
Bij een normale verdeling.
$(\bar{y}-s, \bar{y}+s)$ bevat ongeveer 68% van de waarnemingen.
$(\bar{y}-2s, \bar{y}+2s)$ 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.
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
$P(\bar{A}) = 1 - P(A)$

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.
discrete variabelen

Normale verdeling = $y \sim N(\mu, \sigma)$
Met de GR: ncdf(link.grens, recht.grens, μ, σ)
Standaard normale verdeling: N(0, 1)
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 toevalsvariabele: verwachting $= \mu = E(y) = \sum y_i \cdot p_i = \sum$ uitkomst $\cdot$ kans

variantie van $y = \sigma^2 = Var(y) = \sum (y_i - \mu)^2 \cdot p_i$

Rekenregels voor verwachtingen $\rightarrow \mu_{a+by} = a + b \cdot \mu_y$

$\rightarrow \mu_{x+y} = \mu_x + \mu_y$

voor varianties $\rightarrow \sigma^2_{a+by} = b^2 \cdot \sigma^2_y$

$\rightarrow \sigma^2_{x+y} = \sigma^2_x + \sigma^2_y$

$\rightarrow \sigma^2_{x-y} =$ idem dito

Som van trekkingen uit norm. verdeling

$\sum y \sim N(n \cdot \mu_y, \sqrt{n} \cdot \sigma_y)$

Gemiddelde van trekkingen uit norm. ver.

$\bar{y} \sim N\left(\mu_y, \frac{\sigma_y}{\sqrt{n}}\right)$ of: $\frac{var(y)}{n}$

Centrale Limietstelling

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

Binomiale toets

H_0: $\pi = \ldots$ H_a: $\pi <$ of $>$ of $\neq \ldots$

P-waarde = De kans dat, berekend onder de aanname dat H_0 waar is, de toetsingsgrootheid 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: $\pi \neq \ldots$ doe je een tweezijdige toets. Daarvoor doe je P-waarde $* 2$.

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 Ha juist is.

Normale benadering

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

$y \sim N\left(n\pi, \sqrt{n\pi(1-\pi)}\right) \rightarrow$ 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 $\rightarrow$ Toetsingsgrootheid: $z = \dfrac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$

Als H_0 waar is, is $z \sim N(0, 1)$.

$P(z \geq$ uitkomst TG$) =$ 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) = kans
$Z = (y-\mu)/\sigma$ "z-score van y"
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 = $\Sigma y \rightarrow \Sigma y \sim N(n\mu, \sqrt{n}\sigma)$
} 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 σ_y → op de GR: S_x

$\sigma_{\bar{y}} = \frac{\sigma_y}{\sqrt{n}} = \frac{\sigma_y^2}{n}$ (σ = bekend)
$se(\bar{y}) = \frac{s}{\sqrt{n}}$ = standaardfout van het gemiddelde. (σ = onbekend)

Btbh-i → $\bar{y} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ 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

① H_0: $\mu = \mu_0$ H_a: $\mu <$ of $\neq$ of $> \mu_0$
② TG: $z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$
③ Onder H_0 is $z \sim N(0,1)$ verdeeld.
④ Onder H_a heeft TG de neiging om kleinere/grotere/grotere of kleinere waarden aan te nemen dan onder H_0.
⑤ Linkszijdige/rechtszijdige/tweezijdige toets.
⑥ Reken TG uit.
⑦ $P(z \leq$ uitkomst stap 6)? Voor tweezijdige toets: neem 2x die waarde.
↳ of: $\geq$
⑧ P-waarde $\leq \alpha$ → H_0 verwerpen. H_a is aangetoond.
P-waarde $> \alpha$ → H_0 niet verwerpen. H_a is niet aangetoond.

T-toets

① Hetzelfde als z-toets.
② TG: $t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$
③ Onder H_0 heeft t een t-verdeling met n−1 df.
④ ⑤ ⑥ Hetzelfde als z-toets.
⑦ $P(t_{n-1} \geq$ uitkomst stap 6)
⑧ Hetzelfde als z-toets.

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

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.

2 Steekproeven (EAS)

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

(s_p → samengestelde standaardafwijking; s_p^2 → samengestelde (pooled) variantie)

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

① H_0: $\mu_1 - \mu_2 = 0$ H_a: $\mu_1 - \mu_2 <$ of $\neq$ of > 0 ~~t-toets~~
② TG: $t = \frac{\bar{y}_1 - \bar{y}_2 - 0}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
③ $t_{n_1+n_2-2}$-verdeling

$\boxed{\sigma_1 \neq \sigma_2}$ → $\bar{y}_1 - y_2 \pm t_{\alpha/2} \cdot \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$
↳ uit t-verdeling met df. volgens SPSS-uitvoer.

② TG: $t' = \frac{\bar{y}_1 - \bar{y}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
③ t-verdeling met df. volgens SPSS-uitvoer.

Toets van Levene: kijk in SPSS-uitvoer bij Levene's test → 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 $H_0: \mu_d = D_0 \ (=0)$
② TG: $t = \dfrac{\bar{d} - D_0}{s_d/\sqrt{n}}$
③ t-verdeling met df. = n−1.

$\rightarrow \bar{d} \pm t_{\alpha/2} \cdot \dfrac{s_d}{\sqrt{n}}$
↳ uit 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.
$= \dfrac{1}{n-1} \cdot \sum \left(\dfrac{x-\bar{x}}{s_x}\right)\left(\dfrac{y-\bar{y}}{s_y}\right)$ Schaal-onafhankelijk. Gevoelig voor uitbijters. $-1 \leq R \leq 1$

Regressielijn = $\mu_y = \beta_0 + \beta_1 \cdot x$
of: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

→ β_1 = Verwachte toename in y bij een toename van x van 1 eenheid.
→ β_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 $N \sim (0, \sigma)$ → qq-plot vd. residuen
2) σ is constant → Residuendiagram → geen patroon, dan klopt het.

R^2 = De door de regressie verklaarde fractie y-variatie.
s^2 = Residuele variantie = Mean square Residual (SPSS) = $\dfrac{\sum (\text{residuen})^2}{n-2}$

Btbhi → β_0: $\hat{\beta}_0 \pm t_{\alpha/2} \cdot se(\hat{\beta}_0)$
→ β_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: $t = \dfrac{\hat{\beta}_i - d}{se(\hat{\beta}_i)}$
③ t_{n-2} verdeling.

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

Modelaannames CLS → n moet voldoende groot zijn.
Btbh-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_i zijn onafhankelijk en normaal verdeeld.