STATISTIEK I

R-CODES

R-code	*Betekenis*
Dim ()	Aantal rijen (= aantal personen id steekproef) en kolommen (= aantal variabelen) vd tabel
Head	1e 6 rijen + bijhorende kolommen
Names ()	Namen variabelen
$	Waarde ve variabele uit de date (altijd gevolgd door de naam vd data)
Table ()	Absolute/relatieve frequentie
Read.table ()	Data lezen
Pie ()	Cirkeldiagram
Barplot ()	Staafdiagram
Cut ()	Klassen
Breaks	Grenzen tss de klassen
Hist ()	Histogram
Cumsum ()	Cumulatieve absolute frequentie
Ecdf () *Plot ()*	Cumulatieve frequentiecurve
Mean ()	Gemiddelde
Median ()	Mediaan
Min ()	Minimum
Max ()	Maximum
Aad ()	Gemiddelde absolute afwijking
Var ()	Variantie
Sd ()	Standaarddeviatie Bv. *Sd = sqrt(4):* standaarddeviatie = $\sqrt{4} = 2$
Sqrt ()	Vierkantswortel (vd variantie om de standaarddeviatie te vinden)
Quantile ()	Percentielen/kwartielen
IQR ()	Interkwartielafstand
Boxplot ()	Boxplot
Cov ()	Covariantie
Cor ()	Correlatiecoëfficiënt
Cor () + optie "Kendall"	Kendall's tau

Dbinom (k,N,p)	Kansdichtheid P(X = k)
Pbinom (k,N,p)	Cumulatieve verdelingsfunctie (P ≤ k)
Pnorm (x,μ, σ)	Kans P(X ≤ x) voor variabele $X \sim N(\mu, \sigma^2)$
Pnorm(bekomen Z)	!! in R dus niet σ^2maar σ (door vierkantswortel te nemen) !! bij standaardnormale mogen μ, σ worden weggelaten
Dnorm (x,μ, σ)	Kansdichtheid $f_x(x)$
Pchisq (y,k)	Kansen P(Y ≤ y) voor variabele Y ~ x_k^2
Qchisq(f_Y (y),k)	= y
Pt (t,k)	Kansen P(T ≤ t) voor varibale T ~ t_k
t.test ()	t-toets (g-toets maar in R → t)
mu	H_0
t	Toetsingsgrootheid g
Df()	= degrees of freedom: aantal vrijheidsgraden = k = n -1
p-value	= p-waarde = onderscheidingskans
t.test (alternative = "less")	Linkszijdige toets
t.test (alternative = "greater")	Rechtszijdige toets

Belangrijke commando's

1. q

   = quantile: waarde

   - We weten de kans
   - We willen de waarde weten waarvoor …% ligt eronder/boven

2. p

   = probability: kans

   - We weten de waarde
   - We willen de kans weten die onder/boven een bep waarde ligt

FORMULES EN BIJHORENDE SYMBOLEN

!! steeds finaleresultaat (≠ tussenbewerkingen) afronden op 2 decimalen (<5 naar ↘, ≤ 5 naar ↗)

NAAM	FORMULE	HOE?
RELATIEVE FREQUENTIE	$\frac{Absolute frequentie}{Aantal steekproefelementen}$	**Kennen**
HARMONISCH GEMIDDELDE	$\frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$ x_i = waarde vh i^e element	Zullen we niet echt hanteren
MEETKUNDIG OF GEOMETRISCH GEMIDDELDE	$\sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$	Zullen we niet echt hanteren
(REKENKUNDIG) STEEKPROEFGEMIDDELDE $\bar{x}$ **= GEMIDDELDE O.B.V. WAARDEN VE VARIABELE**	$\frac{(x_1 + x_2 + \ldots + x_n)}{n}$ $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$	Begrijpen, niet reproduceren
GEMIDDELDE O.B.V. EEN FREQUENTIEVERDELING	$\bar{x} = \frac{1}{n} \sum_{i=1}^{p} f_i x_i^u$ Waarbij - x_i^u= unieke waarde vd variabele X in steekproef Bv. x_1^u = vrouw en x_2^u= man - f_i = absolute frequentie van deze waarde - p = aantal unieke waarden vd variabele X id steekproef !! uitkomst hiervan = uitkomst gem. o.b.v. waarden (logisch)	**Kennen**
KLASSENMIDDEN	$\frac{a+b}{2}$!! klassenmiddens]a,b] = [a,b[=]a,b[= [a,b]	Begrijpend lezen

GEMIDDELDE VAN GEGROEPEERDE DATA	$$\bar{x} = \frac{1}{n}\sum_{i=1}^{p} f_i(\frac{a_i + b_i}{2})$$ Waarbij $f_i(\frac{a_i + b_i}{2})$ = absolute frequentie · klassenmidden !! uitkomst hiervan = ± uitkomst gem. o.b.v. waarden of frequentieverdeling	Begrijpend lezen
MEDIAAN md_x	Bij oneven waarden: middelste waarde in **geordende** rij Bij even waarden: middelste 2 waarden gedeeld door 2	**Kennen**
MODUS mo	Waarde/klasse met de hoogste frequentie !! er kunnen meerdere waarden/klassen zijn = **modi** - 1 modus = **unimodaal** - 2 modi = **bimodaal**	**Kennen**
VARIATIEBREEDTE vx	Grootste – kleinste waarde OF Bovengrens laatste klasse – ondergrens eerste klasse (= 2 uitersten) !! kan nooit negatief zijn: max (altijd) > min	**Kennen**
GEMIDDELDE ABSOLUTE AFWIJKING ga_x	$$\frac{\left\|x_1 - \bar{x}\right\| + \left\|x_2 - \bar{x}\right\| + \ldots + \left\|x_n - \bar{x}\right\|}{n}$$ $$ga_x = \frac{1}{n}\sum_{i=1}^{n}\left\|x_i - \bar{x}\right\|$$!! absolute waarde: anders altijd 0 (+ en - heffen elkaar op bij gem.)	**Kennen**
STEEKPROEFVARIANTIE sn_x^2 of s_x^2	$$sn_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$ $$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$	Formularium

STEEKPROEFVARIANTIE O.B.V. FREQUENTIEVERDELING	$$sn_x^2 = \frac{1}{n}\sum_{i=1}^{p} f_i\left(x_i^u - \bar{x}\right)^2$$ $$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} f_i\left(x_i - \bar{x}\right)^2$$ Waarbij - x_i^u= unieke waarde vd variabele X in steekproef Bv. x_1^u = vrouw en x_2^u= man - f_i = absolute frequentie van deze waarde	Formularium **Weten**
STEEKPROEFSTANDAARDDEVIATIE sn_x	Vierkantswortel vd variantie $$sn_x = \sqrt{sn_x^2}$$ $$s_x = \sqrt{s_x^2}$$	**Kennen**

PERCENTIEL P_k	$$\frac{F(P_k)}{n} = \frac{k}{100}$$ Waarbij - P_k = het k-de percentiel - $\frac{F(P_k)}{n}$ = cumulatieve relatieve frequentie Bv. Voor het 10e percentiel: 1) k = 10 2) Cumulatieve relatieve frequentie = k/ 100 = 10% 3) 10% vd waarden zijn hetzelfde of kleiner Bijzondere percentielen: Kwartiel Sym-bool % vd waarden die gelijk zijn of eronder liggen 1e kwartiel P_{25} 25% 2e kwartiel P_{50} 50% 3e kwartiel P_{75} 75% 4e kwartiel P100 100% !! $m\,d_x = P_{50}$	**Kennen**
INTERKWARTIELAFSTAND Q	Verschil 3e en 1e kwartiel $P_{75} - P_{25}$	**Weten**
INTERKWARTIELINTERVAL	$[P_{75}, P_{25}]$!! bevat 50% van alle waarden	**Weten**

SPREIDINGSMAAT D	$$d = \frac{1 - \frac{f_{mo}}{n}}{1 - \frac{1}{p}}$$ Waarbij - p = aantal unieke waarden - f_{mo} = aantal keer dat de m - n = aantal steekproefelementen !! van 0 (geen spreiding) t.e.m. 1 (maximale spreiding)	Formularium
OUTLIERS	$P_{25} - 1{,}5 \cdot Q$ $P_{75} + 1{,}5 \cdot Q$	**Weten**
COVARIANTIE cov_{xy}	$cov_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$	Formularium

CORRELATIECOËFFICIËNT r_{xy}	$$r_{XY} = \frac{cov_{XY}}{s_X s_Y}$$ 1. Covariantie berekenen $$cov_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$ 2. Standaarddeviatie berekenen voor x $$\sqrt{s_x^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$ 3. Standaarddeviatie berekenen voor y $$\sqrt{s_y^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$ 4. $\frac{covariantie}{standaarddeviatie\,x \cdot standaarddevia}$!! correlatiecoëffieciënt heeft altijd zelfde teken als covariantie	**Weten**
CONCORDANT PAAR	$$\frac{y_j - y_i}{x_j - x_i} > 0$$ = positieve hellingsgraad	**Kennen**
DISCORDANT PAAR	$$\frac{y_j - y_i}{x_j - x_i} < 0$$ = negatieve hellingsgraad	**Kennen**
KENDALL'S TAU τ	$\tau = \frac{2(\#concordante\,paren - \#discordante\,paren)}{n(n-1)}$	Formularium
REGRESSIELIJN	$$Y = b_0 + b_1 X$$ Waarbij - b_1 = regressiecoëfficiënt: helling vd rechte - b_0 = intercept: snijpunt met y-as	Formularium

REGRESSIECOËFFICIËNT BIJ PERFECT LINEAIR VERBAND	$$b_1 = \frac{y_j - y_i}{x_j - x_i}$$	**Kennen**
INTERCEPT BIJ PERFECT LINEAIR VERBAND	$$b_0 = y_i - b_1 x_i$$	**Kennen**
REGRESSIECOËFFICIËNT BIJ NIET PERFECT LINEAIR VERBAND	$$b_1 = r_{XY}\frac{s_Y}{s_X}$$!! b_1 zal altijd zelfde teken hebben als r_{xy}	Formularium
INTERCEPT BIJ NIET PERFECT LINEAIR VERBAND	$$b_0 = \bar{y} - b_1\bar{x}$$	Formularium
VANAF HIER: OP POPULATIE NIVEAU		
LIMIET VAN DE RELATIEVE FREQUENTIE (WANNEER STEEKPROEF = ∞)	$$P(X = x_i) = \lim_{n\to\infty}\frac{f_i}{n}$$	Begrijpen, niet reproduceren
CUMULATIEVE VERDELINGSFUNCTIE $F_X(x)$ BIJ DISCRETE EN CONTINUE VARIABELEN	Univariate: $F_X(x) = P(X \le x)$ Bivariate: $F_{X,y}(x,y) = P(X \le x\ en\ Y \le y)$	Begrijpen, niet reproduceren
KANS BIJ CONTINUE VARIABELEN	$$P(X = x) = 0$$!! daarom beroep doen op dichtheidsfunctie	Begrijpen, niet reproduceren
DICHTHEIDSFUNCTIE OF KANSDICHTHEID	$$fx(x) = \lim_{b\to 0}\frac{F_x(x+b) - F_x(x)}{b}$$	Begrijpen, niet reproduceren
INTEGRATIE DICHTHEIDSFUNCTIE	$$P(x_1 \le X \le x_2) = \int_{x_1}^{x_2} fx(x)dx$$	Begrijpen, niet reproduceren
	$$P(X \le x) = \int_{-\infty}^{x} fx(x)dx$$	Begrijpen, niet reproduceren
	$$P(X > x) = \int_{x}^{+\infty} fx(x)dx$$	
KANS BIJ CONTINUE VARIABELEN O.B.V. EIGENSCHAP	$P(x_1 \le X \le x_2) = P(X \le x_2) - P(X \le x_1) = F_x(x_2) - F_x(x_1)$	**Kennen**
VOLLEDIGE OPPERVLAKTE ONDER DICHTHEIDSFUNCTIE	$$\int_{-\infty}^{+\infty} fx(x)dx = 1$$	Begrijpen, niet reproduceren

POPULATIE GEMIDDELDE OF VERWACHTINGSWAARDE $E(X),\ \mu_x\ of\ \mu$	Discreet: $E(X) = \sum_{i=1}^{p} P(X = x_i)x_i$ Continu: $E(X) = \int_{-\infty}^{+\infty} fx(x)dx$	Formularium Begrijpen, niet reproduceren
POPULATIE VARIANTIE $V(X),\ \sigma_x^2 of \sigma^2$	Discreet: $V(X) = \sum_{i=1}^{p} P(X = x_i)(x_i - E(x))^2$ Continu :$V(X) = \int_{-\infty}^{+\infty} fx(x)(x - E(X))^2 dx$	Formularium Begrijpen, niet reproduceren
POPULATIESTANDAARDDEVIATIE $\sigma_x\ of\ \sigma$	$\sigma_x = \sqrt{V(X)}$	**Weten**
UNIVARIATE KANSVERDELING BIJ DISCRETE VARIABELEN	$P(X = x_i) = \sum_{j=1}^{q} P(X = x_1\ en\ Y = y_j)$ $P(Y = y_i) = \sum_{i=1}^{p} P(X = x_1\ en\ Y = y_j)$ Waarbij - p = aantal mogelijke waarden dat X kan aannemen - q = aantal mogelijke waarden dat Y kan aannemen	

STATISTISCHE ONAFHANKELIJKHEID	Discreet: 2 discrete variabelen X en Y zijn onafhankelijk indien: $P\left(X = x_i \text{en} Y = y_j\right) = P\left(X = x_i\right)P\left(Y = y_j\right)$ geldt voor alle mogelijke combinaties i en j 1. Alle kansen neerschrijven 2. Marginale verdeling berekenen 3. Marginale verdeling vermenigvuldigen 4. Deze aan elkaar stellen Continu: 2 continue variabelen X en Y zijn onafhankelijk indien: $P\left(X \leq x \text{en} Y \leq y\right) = P(X \leq x)P\left(Y \leq y\right)$ Voor alle mogelijke waarden x en y	**Kennen** Niet kunnen
POPULATIE COVARIANTIE $\mathbf{COV}(\boldsymbol{X}, \boldsymbol{Y})$	Discreet: $\mathrm{COV}(X, Y) = \sum_{i=1}^{p}\sum_{j=1}^{q} P(X = x_i \, en \, Y = y_j)(y_j - E(Y)$ Continu: $COV(X, Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} fx, y(x, y)(x - E(X))(y - E(Y))dxdy$	Formularium Begrijpen, niet reduceren
POPULATIE CORRELATIECOËFFICIËNT ρ_{XY}	Discreet: $\rho_{XY} = \frac{\mathrm{COV}(X, Y)}{\sigma_X \sigma_Y}$ Continu: $\rho_{XY} = \frac{\mathrm{COV}(X, Y)}{\sigma_X \sigma_Y}$ Waarbij - σ_X = standaarddeviatie van x - σ_Y = standaarddeviatie van y	**Kennen**
EIGENSCHAP POPULATIE GEMIDDELDE OF VERWACHTINGSWAARDE	E(a) = a Indien a = constante Bv. Iedereen 2000 euro inkomen → gem. = 2000	**Weten**

EIGENSCHAP POPULATIE VARIANTIE	V(a) = 0 Bv. Iedereen zelfde inkomen: geen spreiding en variantie = 0	**Weten**
BIJZONDERE VERDELINGEN		
BINOMIALE KANSVERDELING $X \sim Binom(N, p)$	$$P(X = k) = \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k}$$ Waarbij - N = max. aantal successen - k = aantal gewenste successen - p = de kans op een succes - ! = faculteit (via GRM uitrekenen) Bv. 4! = 4·3·2·1 Enkel wanneer: - N = vast - p blijft ongewijzigd Op grafiek - Bij kleine kans op succes: scheef naar rechts Logisch want meeste hebben lage score en enkel de uitzonderingen een hoge - Bij helft kans op succes: symmetrisch - Bij grote kans op succes: scheef naar links Logisch want meeste hebben hoge score en enkel de uitzonderingen een lage	Formularium **Weten**
POPULATIE GEMIDDELDE OF VERWACHTINGSWAARDE VAN EEN BINOMIALE VARIABELE	$$E(X) = N \cdot p$$	Formularium
POPULATIE VARIANTIE VAN EEN BINOMIALE VARIABELE	$$V(X) = N \cdot p\,(1-p)$$	Formularium

DICHTHEIDSFUNCTIE NORMALE VERDELING $X \sim N(\mu, \sigma^2)$	$$fx(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$ Waarbij - π = 3,14… - e = 2,71 - μ = populatie gemiddelde = E(X) - σ^2 = populatie variantie = V(X) - σ = populatie standaarddeviatie Op grafiek: - Hoogste punt (top) = gemiddelde - Grote variantie = laag + breed - Kleine variantie = hoog + smal - Symmetrisch - Enkel positieve waarden	Formularium **Weten** **Weten**
INTEGRAAL VAN DE KANS VAN EEN NORMALE VERDELING	$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$$	Begrijpen, niet reduceren

PRINCIPES STANDAARDNORMALE VERDELING $X \sim N(0,1)$

Weten

$\mu = 0$

$\sigma^2 = 1$

2 belangrijke eigenschappen

1. Symmetrisch rond 0

   Bijgevolg:

   $$P(X > x) = P(X < -x)$$

2. Totale opp. = 1

   Bijgevolg:

   $$P(X \leq -x) = 1 - P(X \leq x)$$

   $$P(X \geq -x) = 1 - P(X \geq x)$$

   $$P(X < x) = 1 - P(X > x)$$

   $$P(X > x) = 1 - P(X < x)$$

   Sidenote:

   - Bij - teken veranderen haakjes van kant
   - < = ≤ en > = ≥ (want continue variabelen)

STANDAARDISEREN VAN X BIJ EEN NORMALE VERDELING

Omdat $\mu \neq$ altijd 0 en $\sigma^2 \neq$ altijd 1 → standaardiseren:

1. $P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right)$ Formularium

Weten

2. $Z = \frac{X-\mu}{\sigma}$ → dit vervangen in bovenstaande formule
3. $P(X \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right)$

Hierdoor: nieuwe variabele Z die de standaardnormale verdeling wél volgt → $Z \sim N(0,1)$

VARIABELE VAN DE CHI-KWADRAAT VERDELING $Y \sim \chi_k^2$	$$Y = x_1^2 + x_2^2 + \ldots + x_k^2$$ Waarbij - k = aantal vrijheidsgraden (en ook populatie gemiddelde) - $$\chi_k \sim N(0,1)$$ Op grafiek: - Hoogste punt (top) = k en dus ook populatie gemiddelde - Scheve verdeling: asymmetrisch - Enkel positieve waarden - Totale opp./kans = 1	Formularium **Weten**
POPULATIE GEMIDDELDE BIJ EEN CHI-KWARDRAAT VERDELING	$$E(Y) = k$$	Formularium
POPULATIE VARIANTIE BIJ EEN CHI-KWARDRAAT VERDELING	$$V(Y) = 2k$$	Formularium
VARIABELE VAN DE STUDENT-T VERDELING $T \sim t_k$	$$T = \frac{X}{\sqrt{\frac{1}{k}Y}}$$ k = aantal vrijheidsgraden Op grafiek: - Indien k → ∞(= t_∞): valt exact samen met standaardnormale - Hoogste punt (top) = populatie gemiddelde = 0 - Symmetrisch	Formularium Interpretatie ervan niet kennen **Weten**
POPULATIEGEMIDDELDE BIJ EEN T-VERDELING	$$E(T) = 0$$	Formularium
VARIANTIE BIJ EEN T-VERDELING	$$V(T) = \frac{k}{k-2},\ voor\ k > 2$$	Formularium

VANAF HIER: STEEKPROEVENVERDELING

Steekproefgemiddelde

STEEKPROEFGEMIDDELDE $\bar{X}$ **= GEMIDDELDE VAN VERSCHILLENDE STEEKPROEVEN**	$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$!! $\bar{X}$ = steekproefgemiddelde voor een bepaalde steekproef, voor een steekproef ih algemeen $\bar{x}$ = steekproefgemiddelde o.b.v. 1 specifieke steekproef	**Kennen**
VERWACHTINGSWAARDE VAN HET STEEKPROEFGEMIDDELDE $E(\bar{X})$	$$E(\bar{X}) = \mu_x$$	Formularium
VARIANTIE VAN HET STEEKPROEFGEMIDDELDE $V(\bar{X})$	$$V(\bar{X}) = \frac{\sigma_x^2}{n}$$	Formularium
VERDELING VAN HET STEEKPROEFGEMIDDELDE **= STEEKPROEFVERDELING VAN HET GEMIDDELDE**	Bij onafhankelijke, lukrake trekkingen uit populatie dat normaal verdeeld is: $$\bar{X} \sim N(\mu_x, \frac{\sigma_x^2}{n})$$	Formularium
STANDAARDISEREN VAN HET STEEKPROEFGEMIDDELDE	$$Z \leq \frac{x-\mu}{\sqrt{\sigma_x^2 / n}}$$ $$\text{of } Z \leq \frac{x-\mu}{\sigma_x/\sqrt{n}}$$ Voorwaarde: - X komt uit normale verdeling Hierbij: n maakt niet uit OF - n ≥ 30	**Kennen**
Steekproefvariantie		
STEEKPROEFVARIANTIE SN_x^2 of S_x^2	$$SN_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$ $$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$	**Kennen**

VERWACHTINGSWAARDE VOOR STEEKPROEFVARIANTIE	$$E(SN_x^2) = \frac{n-1}{n}\sigma_x^2$$ $$E(S_x^2) = \sigma_x^2$$!! daarom: in praktijk meer S_x^2 dan SN_x^2	Formularium
VERDELING VAN DE STEEKPROEFVARIANTIE	Bij onafhankelijke, lukrake trekkingen uit populatie dat normaal verdeeld is: $$\frac{(n-1)S_x^2}{\sigma_x^2} \sim \chi_{n-1}^2$$	Formularium
SCHATTERS		
EEN GOEDE SCHATTER $\hat{\theta}$ VOOR EEN POPULATIEPARAMETER θ	$\hat{\theta}$ is een goede schatter voor θ indien: 1. De schatter zuiver is: verwachtingswaarde schatter = populatieparameter $$E\left(\hat{\theta}\right) = \theta$$ 2. De variantie vd schatter $V\left(\hat{\theta}\right)$ kleiner wordt naarmate de steekproefgrootte n ↗ = naarmate n ↗ wordt de schatter nauwkeuriger	**Weten**
STANDAARDDEVIATIE VAN DE SCHATTER = STANDAARDFOUT	$$\sqrt{V\left(\hat{\theta}\right)}$$ Of $\frac{\sigma}{\sqrt{n}}$ want $V(\bar{X}) = \frac{\sigma_x^2}{n}$ Schatter met kleinste standaardfout = het efficiënst	**Weten**

STEEKPROEFGEMIDDELDE: EEN GOEDE SCHATTER VOOR POPULATIEGEMIDDELDE?

Steekproefgemiddelde = goede schatter want **Weten**

1. $$E(\bar{X}) = \mu_x$$

   Dus gem. van alle steekproefgem. ≈ populatie gem.

2. Variantie steekproefgemiddelde = $$V(\bar{X}) = \frac{\sigma_x^2}{n}$$

   Dus naarmate n ↗ → nauwkeuriger

⇒ Ja!! goede schatter

STEEKPROEFVARIANTIE: EEN GOEDE SCHATTER VOOR POPULATIEVARIANTIE?

2 formules voor steekproefvariantie → bekijken beide **Weten**

<u>Formule 1:</u> $sn_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Deze is geen goede schatter, want

1. $$E(SN_x^2) = \frac{n-1}{n}\sigma_x^2$$

   Dus σ_x^2 (populatievariantie) zal steeds te klein worden geschat (door (n-1)/n

2. Naarmate n ↗ zal het wel nauwkeuriger worden

⇒ slechts 1 vd 2 voorwaarden is voldaan

⇒ geen goede schatter

<u>Formule 2:</u> $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

1. $$E(S_x^2) = \sigma_x^2$$

   Verwachting vd steekproefvariantie = populatievariantie → goed!!

2. Naarmate n ↗ zal het wel nauwkeuriger worden

⇒ beide voorwaarden zijn voldaan

⇒ deze formule = voorkeur

2 MOGELIJKHEDEN OM O.B.V. SCHATTING UITSPRAAK TE DOEN OVER POPULATIEPARAMETER

Via betrouwbaarheidsinterval (enkel tweezijdig kunnen)

BETROUWBAARHEIDSINTERVAL WANNEER VARIANTIE GEKEND	$\left[\bar{X} - z_{\frac{\alpha}{2}}\sigma/\sqrt{n}, \bar{X} + z_{\frac{\alpha}{2}}\sigma/\sqrt{n}\right]$ Z_α = waarde vd standaardnormale verdeling zodat de opp. vd curve <u>rechts</u> vd waarde = α	**Kennen**
BETROUWBAARHEIDSINTERVAL WANNEER VARIANTIE ONGEKEND IS	$\left[\bar{X} - t_{n-1;\alpha/2}S_x/\sqrt{n}, \bar{X} + t_{n-1;\alpha/2}S_x/\sqrt{n}\right]$ $T_{n-1;\alpha/2}$ = waarde vd t-verdeling zodat de opp. vd curve <u>rechts</u> vd waarde = α	Formularium

Via statistisch toetsen (eenzijdig en tweezijdig kunnen)

(Hierbij gaan we er altijd vanuit dat x normaal verdeeld is of n ≥ 30 is)

NULHYPOTHESE	Bij tweezijdig, linkszijdig en rechtszijdig $H_0: \mu = \mu_0$ μ_0 = gegeven waarde	Formularium
ALTERNATIEVE HYPOTHESE	Tweezijdig: $H_a: \mu \neq \mu_0$ Linkszijdig: $H_a: \mu < \mu_0$ Rechtszijdig: $H_a: \mu > \mu_0$	Formularium

TOETSINGSGROOTHEID

$$G = \frac{\bar{X} - \mu_0}{S_X / \sqrt{n}}$$

Formularium

Wanneer nulhypothese waar is ($\mu = \mu_0$):

Weten

- G volgt t_{n-1}-verdeling
- Waarden van G liggen rond 0
- Waarden van G zijn + en -

Wanneer nulhypothese niet waar is en $\mu > \mu_0$:

- Grotere waarden $\leftrightarrow \mu = \mu_0$
- G heeft enkel positieve waarden

Wanneer nulhypothese niet waar is en $\mu < \mu_0$:

- Kleinere waarden $\leftrightarrow \mu = \mu_0$
- G heeft enkel negatieve waarden

g = waarde van G die we bekomen o.b.v. één steekproef

Maar: wanneer wat doen met g-waarde?

- Als g rond 0 ligt → H_0 niet verwerpen
- Als g sterk van 0 verschilt → H_0 wel verwerpen

Wat is “rond” en wat is “sterk verschilt”? → **verschillende beslissingsregels**

1. **BESLISSINGSREGELS O.B.V. AANVAARDINGSGEBIED/KRITIEKE WAARDEN** Komen overeen met interval vd tweezijdige toets	Tweezijdig: Indien g tussen het volgende interval ligt: $$-t_{n-1;\alpha/2} \leq g \leq t_{n-1;\alpha/2}$$ H_0 <u>niet</u> verwerpen Ligt het er niet in → H_0 wél verwerpen M.a.w.: $\lvert g \rvert \leq t_{n-1;\alpha/2}$ → H_0 <u>niet</u> verwerpen $\lvert g \rvert > t_{n-1;\alpha/2}$ → H_0 verwerpen, H_a aanvaarden Linkszijdig: $g > -t_{n-1;\alpha}$ → H_0 <u>niet</u> verwerpen $g < -t_{n-1;\alpha}$ → H_0 verwerpen, H_a aanvaarden Rechtszijdig: $g \leq t_{n-1;\alpha}$ → H_0 <u>niet</u> verwerpen $g > t_{n-1;\alpha}$ → H_0 verwerpen, H_a aanvaarden α = significantie niveau	Formularium **Weten** Formularium **Weten** Formularium
TYPE I FOUT	H_0 = correct maar we verwerpen ze → **type I fout**	**Weten**
KANS OP EEN TYPE I FOUT	P(verwerp H_0 | $\mu = \mu_0$| = α α = het significantieniveau en is altijd <u>gegeven</u>	**Weten**

KANS OP CORRECT BESLUIT INDIEN $\mu = \mu_0$	P(verwerp H_0 niet$\mid \mu = \mu_0 \mid$= 1 - α OF Kijken naar …% betrouwbaarheidsinterval Bv. Bij 95% betrouwbaarheidsinterval: 1. α = 0,05 1-0,05 = 0,95 Dit: kans op correct besluit OF 2. 95% betrouwbaarheidsinterval dus 95% kans op correct besluit	**Weten**
TYPE II FOUT	H_0 ≠ correct maar we verwerpen ze niet → **type II fout**	**Weten**
KANS OP EEN TYPE II FOUT	P(verwerp H_0 niet $\mid \mu \neq \mu_0 \mid$= β β = bèta	**Weten**
KANS OP CORRECT BESLUIT INDIEN $\mu \neq \mu_0$ **= ONDERSCHEIDINGSKANS OF POWER**	P(verwerp $H_0 \mid \mu \neq \mu_0 \mid$= 1 - β Invloeden - Significantieniveau Wanneer $\alpha \nearrow \rightarrow \beta \searrow$ - Steekproefgrootte Wanneer n $\nearrow \rightarrow \beta \searrow$ En dus kans op correct besluit (1 - β) $\nearrow$	**Weten**
2. **BESLISSINGSREGELS O.B.V. BETROUWBAARHEIDSINTERVAL**	Indien μ_0 tussen het volgende interval ligt: $\left[\bar{X} - t_{n-1;\alpha/2} S_x/\sqrt{n}, \bar{X} + t_{n-1;\alpha/2} S_x/\sqrt{n}\right]$ H_0 niet verwerpen Ligt het er niet in → H_0 wél verwerpen	Formularium

3.

BESLISSINGSREGELS O.B.V. P-WAARDE OF OVERSCHRIJDINGSKANS

p-waarde

Formularium

- Wordt berekend in veronderstelling dat H_0 waar is
- Hangt af vd H_a

Basisregel:

$p \geq \alpha \rightarrow H_0$ niet verwerpen

$p < \alpha \rightarrow H_0$ verwerpen

Linkszijdig: $P(G < g \mid \mu = \mu_0)$

1. $P(T < g)$ berekenen
2. Deze g: in R output "pt()" steken
3. Bekomen waarde vergelijken met α + basisregel toepassen

Rechtszijdig:$P(G > g \mid \mu = \mu_0)$

1. $P(T > g)$ berekenen
2. 1 - pt(g,(n-1)) doen
3. Bekomen waarde vergelijken met α + basisregel toepassen

Tweezijdig:

1. g waarde berekenen
2. g waarde vergelijken met 0
   - Als g > 0:

     $p = 2 \cdot P(T > g)$
   - Als g ≤ 0:

     $p = 2 \cdot P(T < g)$

   Deze kansen → aflezen in R

VERDUIDELIJKING POPULATIEPARAMETERS

	Bij discrete variabelen	**Bij continue variabelen**
Populatie gemiddelde / verwachtingswaarde $E(X)$, μ_x **of** μ	$E(X) = \sum_{i=1}^{p} P(X = x_i)x_i$	$P(X = x_i) = 0$ → andere definitie nodig: $E(X) = \int_{-\infty}^{+\infty} f x(x) d x$
Populatie variantie $V(X)$, σ_x^2 **of** σ^2	$V(X) = \sum_{i=1}^{p} P(X = x_i)(x_i - E(x))^2$	$V(X) = \int_{-\infty}^{+\infty} f x(x)(x - E(X))^2 d x$
Populatie covariantie	$\mathrm{COV}(X,Y) = \sum_{i=1}^{p}\sum_{j=1}^{q} P(X = x_i \text{ en } Y = y_j)(y_j - E(Y))$	$COV(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f x,y(x,y)(x - E(X))(y - E(Y)) d x d y$
Populatie correlatiecoëfficiënt		$\rho_{XY} = \frac{\mathrm{COV}(X,Y)}{\sigma_X \sigma_Y}$
	Formularium (behalve correlatie)	Moeten we niet kunnen uitrekenen

STAPPENPLAN BETROUWBAARHEIDSINTERVAL

<u>Wanneer?</u> Populatie gem. is ongekend maar we willen er toch uitspraak over doen

Vraag 1: X normaal verdeeld?		
JA		**NEE**
Vraag 2: Populatievariantie gekend?		
JA	**NEE**	**NEE**

$\left[\bar{X} - z_{\frac{\alpha}{2}}\sigma/\sqrt{n}, \bar{X} + z_{\frac{\alpha}{2}}\sigma/\sqrt{n}\right]$ Ligt het tss dit interval → interval bevat pop.gem. Bv. Bij 95% betrouwbaarheidsinterval → in 95% vd gevallen ligt pop. gem. erin <u>Invloeden</u> 1. Steekproefgrootte Naarmate n ↗ → interval smaller 2. Als α ↗ → interval smaller Want 1 - α zal ↘ en hiertss = interval Indien je de kans wil ↗ dat pop.gem. erin ligt → interval moet breed zijn <u>maar</u> indien breed: niet zo informatief meer ⇒ compromis tss beide vinden (vaak 95%)	$\left[\bar{X} - t_{n-1;\alpha/2}S_x/\sqrt{n},\ \bar{X} + t_{n-1;\alpha/2}S_x/\sqrt{n}\right]$ Nu: t-verdeling ipv normale verdeling Deze: lijken op elkaar Toch verschillen: - T_{n-1}-verdeling heeft een grotere variantie - $T_{n-1;\alpha/2}$-waarde is groter dan $z_{\frac{\alpha}{2}}$ -waarde Maar ook hierbij: naarmate n ↗ → steeds betere benadering standaardnormale verdeling	Indien grote steekproef: centrale limietstelling Indien kleine steekproef: GEEN LEERSTOF VAN STATISTIEK I

STAPPENPLAN STATISTISCH TOETSEN

1. H_0 en H_a opstellen (uit gekregen opgave)
2. Significantieniveau vaststellen
3. Gem. en standaarddevidatie vd specifieke steekproef berekenen (of uit opgave halen)
4. Toetsingsgrootheid g berekenen

$$G = \frac{\bar{X} - \mu_0}{S_X / \sqrt{n}}$$

5. Beslissingsregels toepassen
   - M.b.v. kritieke waarde

     !! bij kritische waarde: 1-P(T < t-waarde) doen

     Dit = P(T > t-waarde) wat altijd zo is (want t-waarde = waarde rechts vd grafiek)
   - M.b.v. betrouwbaarheidsinterval
   - M.b.v. p-waarde

   !! andere beslissingsregels voor eenzijdig, linkszijdig en rechtszijdig

6. Conclusie formuleren: H_0 verwerpen of niet?

- Niet verwerpen = H_0 aanvaarden
- Verwerpen = H_a aanvaarden

Opletten !!

Voor zowel de Z_α - waarde, de $T_{n-1;\alpha/2}$-waarde als de kritische waarde: gaat het om de waarde rechts vd grafiek

Maar: R-output geeft standaard wat links onder grafiek ligt

Oplossing?

1. Eigenschap normale/t-verdeling gebruiken

   1 - … doen

OF

2. Bij R-output extra info geven: “lower.tail”

Voorbeeld

α = 0,025 $\Rightarrow P\,(Z > Z_{0,025}) = \ 0{,}025$

Dus 1 - 0,025 = 0,975

Dit in R-output: $q_{norm}(0{,}975)$

Uitkomst van deze = juiste z-waarde