Naam	Formule	R-code	Extra
Populatieparameters toetsen			
Onbekende verwachting μ en onbekende populatie-variantie	$\frac{\bar{x} - \mu_0}{s / \sqrt{n}} \sim t_{n-1}$	**p-waarde:** t.test(x = vector, mu = …, alternative = "two.sided/less/greater", conf.level = 0.95) **power:** verschil tussen 2 gemiddelden in een steekproef detecteren power.t.test(n/power = …, delta = μ0-μa, sd = …, alternative = "two.sided/less/greater", sig.level = 0.05, type = "one.sample")	Voorwaarden: - X is tenminste van intervalniveau - X is normaal verdeeld of de steekproef is groot
Onafhankelijke onbekende verwachtingen μ1 en μ2 en onbekende populatie-varianties die niet gelijk zijn aan elkaar	**p-waarde:** Welch t-toets: $\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{n}}} \sim t_l$ $l = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_1}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_1-1)}}$ **power:** Effectsize d: $d = \frac{\mu_1 - \mu_2}{\sqrt{s^2_{pooled}}}$ $s^2 pooled = \frac{SS_X + SS_Y}{n_1 + n_2 - 2}$	**p-waarde:** t.test(x = vector, y = vector, alternative = "two.sided/less/greater", conf.level = 0.95) **power:** verschil tussen 2 gemiddelden gedeeld door de gepoolde sd detecteren pwr.t2n.test(n1/power = …, n2/power = …, d = effectsize, sig.level = 0.05, alternative = "less/greater/two.sided") deze functie kan enkel gebruikt worden indien n1 = n2 (en onafhankelijk): power.t.test(delta = μ1-μ2, sd = effectsize d, n/power = …, sig.level = 0.05, alternative = "two.sided/one.sided", type = "two.sample") **normaliteit van 2 onafhankelijke gemiddeldes nagaan**: qqnorm(x1) qqnorm(x2)	Voorwaarden: - X is tenminste van intervalniveau - X is normaal verdeeld in beide populaties, of beide steekproeven zijn groot - De steekproeven zijn onafhankelijk Benodigde package om power te berekenen: "pwr" d is negatief indien μ1-μ2 negatief is Equivalent aan lineaire regressie indien de predictor dichotoom is ➔ Bij Welch: 2 vectoren creëren uit de twee levels van de dichotome variabele → de gemiddelden van deze 2 vectoren ga je vergelijken
Afhankelijke onbekende verwachtingen μ1 en μ2 en onbekende populatie-varianties	$d = \mu_1 - \mu_2$ $\frac{\bar{d}}{s / \sqrt{n}} \sim t_{n-1}$	**p-waarde:** 1 steekproef van maken t.test(x = d, mu = 0, alternative = "greater/less/two.sided") t.test(x = vector, y = vector, alternative = "two.sided/less/greater", paired = TRUE) **power:** power.t.test(n/power = …, delta = μ1-μa, sd = sd van delta, type = "paired", alternative "two.sided/one.sided")	Voorwaarden: - X is tenminste van intervalniveau - X1-X2 moet normaal verdeeld zijn of de steekproef moet groot zijn Bij afhankelijke steekproeven zijn de steekproefgroottes altijd even groot

		normaliteit van 2 afhankelijke gemiddeldes nagaan: 1 steekproef van maken qqnorm(x1 – x2)	
Onbekende proportie	$$P(B(n,\pi) < k) = pwaarde$$ $$\frac{n!}{k!\,(n-k)!}\pi^k(1-\pi)^{(n-k)} \sim B(n,\pi)$$	**p-waarde:** binom.test(x = k, n = n, p = π, alternative = "one.sided/two.sided") **power**: powerBinom(n/power = …, p0 = …, p1 = …, sig.level = 0.5, alternative = "one.sided/two.sided")	Benodigde package om power te berekenen: "exactci" Als Ha tweezijdig is, dan moeten uitgerekende kansen vermenigvuldigd met 2 worden
Onbekende B1	**p-waarde bij enkelvoudig lineair verband:** <u>Via t-verdeling:</u> $$\frac{B_1}{\sqrt{\frac{SS_{Res}}{(n-2)SS_X}}} \sim t_{n-2}$$ <u>Via F-verdeling</u>: nulmodel en lineair model vergelijken $$\frac{SS_{Res0} - SS_{Res1}}{SS_{Res1}/(n-2)} \sim F_{1,n-2}$$ Waarbij SSRes0 gelijk is aan SSY **p-waarde bij meervoudig lineair verband:** <u>Via F-verdeling</u>: model A en model B vergelijken $$\frac{(SS_{ResA} - SS_{ResB})/(df_A - df_B)}{SS_{ResB}/df_B}$$ $$\sim F\left(\frac{(df_A - df_B)}{df_B}\right)$$ Waarbij: df_A: aantal vrijheidsgraden van model A (n-k-1) df_B: aantal vrijheidsgraden van model B (n-p-1) k: subset van predictoren p **power bij enkelvoudig lineair verband**: correlatie schatten om te gebruiken voor R-functie $$\rho = B_1\frac{\sigma_x}{\sigma_y}$$	**p-waarde bij enkelvoudig lineair verband**: <u>Via t-verdeling:</u> pt(q = …, df = …, lower.tail = FALSE) pf(q = …, df1 = 1, df 2 = n-2, lower.tail = FALSE) <u>Via F-verdeling</u>: toetst nulmodel vs lineair model summary(lm): nagaan of een specifieke coëfficiënt 0 is <u>Voor nominale predictor met dummy/effect codering</u>: enkel F-toets interpreteren, t-toets is betekenisloos (t-toetsen zijn wél betekenisvol bij dichotome rpedictoren!!!) summary(lm) **p-waarde bij meervoudig lineair verband**: <u>Via F-verdeling</u>: model A (k) en model B (p) vergelijken anova(lmA, lmB) <u>Via F-verdeling</u>: nulmodel en lineair model vergelijken summary(lm): nagaan of de regressiecoëfficiënten van alle predictoren 0 zijn nulmodel creëren om te gebruiken in anova-functie: lm0 <- lm(formula = y ~ NULL) → anova(lm0, lm) <u>Voor 1 nominale predictor met dummy/effect codering</u>: *enkel F-toets interpreteren!!* summary(lm): nulmodel vs lm, coëfficiënten en gemiddeldes te weten komen	Voorwaarden: - X is tenminste interval, ratio of dichotoom - Y is tenminste interval of ratio - $E(\varepsilon_i) = 0$ voor alle i - $V(\varepsilon_i) = V(\varepsilon_j)$ voor alle i,j - $COV(\varepsilon_i, \varepsilon_j) = 0$ voor alle i,j - ε_i moet normaal verdeeld zijn of de steekproef moet groot zijn F-verdeeld met $df_0 - df_1$ vrijheidsgraden in de teller en df_1 in de noemer F-toets is altijd eenzijdig: **alfa niet delen door 2**! Ongeacht de alternatieve hypothese Package inlezen om power te berekenen: "pwr" Volgorde van de argumenten in de functie anova: eerst het model met minder predictoren (LM), dan het model met meer predictoren (LM.geslacht) Opgelet voor collineariteit: package "car" - VIFs = 1 → goed - VIFs < 3 → oké - VIFs < 10 en één VIF van < 3 → grijze zone - Één VIF > 10 → geen regressie

	power bij meervoudig lineair verband: via F-verdeling het verschil tussen de determinatiecoëfficiënten (voor lm0 vs lm of lmA vs lmB) $$f^2 = \frac{R_B^2 - R_A^2}{1 - R_B^2} \sim F_{p-k\,;\,n-p-k}$$	anova(lm, lmdummy): model zonder en model met dummycodering (die volgorde!), f-toets die nagaat of de extra variantie die het lm met dummy verklaart toevallig is **power bij enkelvoudig lineair verband**: correlatiecoëfficiënt detecteren pwr.r.test(n/power = …, r = ρ, sig.level = 0.05) **power bij meervoudig lineair verband**: verschil van de determinatiecoëfficiënten van de 2 modellen detecteren (voor lm0 vs lm én lmA vs lmB) pwr.f2.test(u = p-k, f2 = f², power = …, sig.level = 0.05) waarbij v = n-p-1 (gebruiken om n te berekenen) **predicties**: fitted(lm) **residuen**: residuals(lm) **SSMod**: sum((fitted(lm)-mean(Y))^2) **SSTot**: var(y)*(n-1)	Bij df_B bij dummycodering: elke hulpveranderlijke wordt gezien als een aparte predictor, en b0 is geen predictor B1 voor Ha berekenen: als Y stijgt met 1, dan stijgt X met 4 = ¼ Bij alle soorten modelvergelijkingen of variatievergelijkingen: het model met het MINST predictoren wordt eerst in de formule/functie gestopt, want dit model heeft de grootste variantie (anders negatief) Bij F- en chikwadraat verdelingen zoeken we altijd kansen in de vorm van P(X>x) (omdat we geïnteresseerd zijn in extreme waarden, en beide verdelingen beginnen pas vanaf 0) Equivalent aan Welch t-toets indien de predictor dichotoom is
Pearson chikwadraat toets Zijn de p proporties in de k populaties identiek? Zijn twee categorische variabelen afhankelijk of niet?	$$X^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} \frac{\left(f_{i,j} - n_i\hat{\pi}_{.j}\right)^2}{n_i\hat{\pi}_{.j}}$$ $$\sim \chi^2_{(p-1)(k-1)}$$ waarbij: p = aantal categorieën van de categorische variabele k = aantal populaties $\hat{\pi}_{.j}$ = schatting van de verwachte proporties in de categorie j indien de nulhypothese correct is	**p-waarde en afhankelijkheid:** chisq.test(table(variabele 1, variabele 2)) **power**: power.chisq.test(w = effectgrootte, N = n, df = (p-1)(k-1), sig.level = 0.05) **effectgrootte w**: ES.w2(bivariate kansverdeling Ha)	**Altijd een tweezijdige toets: zijn de proporties identiek in de twee populaties?** Noemt ook de homogeniteitstoets **De vraag of twee categorische variabelen identiek zijn is equivalent aan het vragen of <u>twee variabelen afhankelijk zijn</u>** Benodigde package om de effectgrootte en de power te berekenen: “pwr” Bij F- en chikwadraat verdelingen zoeken we altijd kansen in de vorm van P(X>x) (omdat we

			geïnteresseerd zijn in extreme waarden, en beide verdelingen beginnen pas vanaf 0)
Puntschattingen (incorrecte geheugensteuntjes)			
Gemiddelde of verwachting	$E(X) = \bar{x}$ $SE_{E(X)} = \frac{\sigma}{\sqrt{n}}$	mean(…)	Voorwaarden: - X is tenminste interval, ratio of absoluut
Variantie	$V(X) = s_x^2$ $E(S_x^2) = \sigma_x^2$ $E(SN_x^2) = \frac{n-1}{n}\sigma_x^2$ $V(X) = \frac{SS_{Res}}{n-1}$	var(…) sd(…)	Voorwaarden: - X is tenminste interval, ratio of absoluut
Covariantie	$COV_{X,Y} = cov_{x,y}$	cov(…, …)	Voorwaarden: - X is tenminste interval, ratio of absoluut
Correlatiecoëfficiënt	$\rho_{X,Y} = r_{x,y} = \frac{cov_x}{s_x s_y}$	cor(…, …)	Voorwaarden: - X is tenminste interval, ratio of absoluut
B1	**Voor een enkelvoudig en meervoudig lineair model**: $\beta_1 = b_1$ $E(B_1) = \beta_1$ **Voor een enkelvoudig lineair model**: $V(B_1) = \frac{\sigma_\varepsilon^2}{SS_X} = \frac{\sigma_\varepsilon^2}{(n-1)s_x^2}$	lm(formula = y ~ x) summary(lm)	Voorwaarden: - X is tenminste interval, ratio of absoluut - $E(\varepsilon_i) = 0$ voor alle i - $V(\varepsilon_i) = V(\varepsilon_j)$ voor alle i,j - $COV(\varepsilon_i, \varepsilon_j) = 0$ voor alle i,j Variantie van B1 verkleinen: - n zo groot mogelijk - s_x^2 zo groot mogelijk - σ_ε^2 zo klein mogelijk Opgelet voor collineariteit: package “car” - VIFs = 1 → goed - VIFs < 3 → oké - VIFs < 10 en één VIF van < 3 → grijze zone - Één VIF > 10 → geen regressie
B0	Voor een enkelvoudig en meervoudig lineair model: $\beta_0 = b_0$	lm(formula = y ~ x) lm(formula = y ~ x1 + x2 + …)	Voorwaarden: - X is tenminste interval, ratio of absoluut - $E(\varepsilon_i) = 0$ voor alle i

	$E(B_0) = \beta_0$ Voor een enkelvoudig lineair model: $V(B_0) = \sigma_\varepsilon^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)$ $= \sigma_\varepsilon^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_X}\right)$	coef(lm) summary(lm)	- $V(\varepsilon_i) = V(\varepsilon_j)$ voor alle i,j - $COV(\varepsilon_i, \varepsilon_j) = 0$ voor alle i,j Variantie van B0 verkleinen: - n zo groot mogelijk - s_x^2 zo groot mogelijk - σ_ε^2 zo klein mogelijk
Predictie Yi	Voor een enkelvoudig en meervoudig lineair model: $\hat{Y}_i = b_0 + b_1 x_i$ $E(Yi\|Xi) = b_0 + b_1 x_i$ Voor een enkelvoudig lineair model: $V(\hat{Y}_i) = \sigma_\varepsilon^2\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}\right)$ $= \sigma_\varepsilon^2\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{SS_X}\right)$		Voorwaarden: - X is tenminste interval, ratio of absoluut - $E(\varepsilon_i) = 0$ voor alle i - $V(\varepsilon_i) = V(\varepsilon_j)$ voor alle i,j - $COV(\varepsilon_i, \varepsilon_j) = 0$ voor alle i,j De variantie van de predictie van Yi stijgt naarmate de predictie verder van het gemiddelde ligt → **predicties zijn beter voor punten die dichter bij het gemiddelde liggen** Variantie van Yi verkleinen: - n zo groot mogelijk - s_x^2 zo groot mogelijk - σ_ε^2 zo klein mogelijk
Variantie van de foutterm (residuen) σ_ε^2	**Voor een enkelvoudig lineair model:** $\sigma_\varepsilon^2 = s_\varepsilon^2 = \frac{1}{n-2}\sum_i^n (y_i - \hat{y}_i)^2 = \frac{SS_{Res}}{n-2}$ **Voor een meervoudig lineair model:** $S_\varepsilon^2 = \hat{\sigma}_\varepsilon^2 = \frac{1}{n-p-1}\sum_{i=1}^n (Y_i - B_0 - B_1 x_{i1} - \cdots - B_p x_{ip})^2$ $= \frac{1}{n-p-1}\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ $= \frac{SS_{Res}}{n-p-1}$	((sum(residuals(lm)^2))/n-1)	Voorwaarden: - X is tenminste interval, ratio of absoluut - $E(\varepsilon_i) = 0$ voor alle i - $V(\varepsilon_i) = V(\varepsilon_j)$ voor alle i,j - $COV(\varepsilon_i, \varepsilon_j) = 0$ voor alle i,j
Betrouwbaarheidsintervallen			
BI voor het gemiddelde	$\left[\bar{x} \pm t_{n-1;\frac{\alpha}{2}} \frac{Sx}{\sqrt{n}}\right] \sim t_{n-1}$	qt(p = 0.025, df = …, lower.tail = FALSE) qt(p = 0.975, df = …)	Voorwaarden: - X is normaal verdeeld of de steekproef is groot

	Breedte: $2\left(t_{n-1;\frac{\alpha}{2}}\frac{Sx}{\sqrt{n}}\right)$		- X is tenminste van intervalniveau Kritieke t-waarde moet **altijd positief zijn**, dus kijken naar de rechterkant van de verdeling Breedte: - Hoe groter sigma, hoe breder - Hoe groter n, hoe smaller - Hoe groter alfa, hoe smaller - Hoe kleiner de kritieke waarde, hoe breder
BI voor b1	Voor enkelvoudige lineaire regressie: $\left[b_1 \pm t_{n-2;\alpha/2}\sqrt{\widehat{V(B_1)}}\right]$ $= \left[b_1 \pm t_{n-2;\alpha/2}\sqrt{\frac{\widehat{\sigma_\varepsilon^2}}{(n-2)\,s_x^2}}\right]$ $= \left[b_1 \pm t_{n-2;\alpha/2}\sqrt{\frac{\widehat{\sigma_\varepsilon^2}}{SS_X}}\right]$ Waarbij $\widehat{\sigma_\varepsilon^2} = \frac{SS_{Res}}{n-p-1}$	**Voor enkelvoudige en meervoudige lineaire regressie:** confint(lm, level = 0.95)	Voorwaarden: - X is normaal verdeeld of de steekproef is groot - X is tenminste van intervalniveau - ε_i is normaal verdeeld - $E(\varepsilon_i) = 0$ voor alle i - $V(\varepsilon_i) = V(\varepsilon_j)$ voor alle i,j - $COV(\varepsilon_i, \varepsilon_j) = 0$ voor alle i,j Smaller BI: - n zo groot mogelijk - s_x^2 zo groot mogelijk - σ_ε^2 zo klein mogelijk
BI voor b0	Voor enkelvoudige lineaire regressie: $\left[b_0 \pm t_{n-2;\alpha/2}\sqrt{\widehat{V(B_0)}}\right]$ $= \left[b_0 \pm t_{n-2;\alpha/2}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{SS_X}}\right]$	**Voor enkelvoudige en meervoudige lineaire regressie:** confint(lm, level = 0.95)	Voorwaarden: - X is normaal verdeeld of de steekproef is groot - X is tenminste van intervalniveau - ε_i is normaal verdeeld - $E(\varepsilon_i) = 0$ voor alle i - $V(\varepsilon_i) = V(\varepsilon_j)$ voor alle i,j - $COV(\varepsilon_i, \varepsilon_j) = 0$ voor alle i,j Smaller BI: - n zo groot mogelijk - s_x^2 zo groot mogelijk - σ_ε^2 zo klein mogelijk
Andere			
Determinatie-coëfficiënt R^2	**Voor enkelvoudige en meervoudige lineaire regressie**:	summary(lm)	Eigenschappen van R^2: - R^2 is nooit negatief

	$$R^2 = \frac{SS_{Mod}}{SS_{Tot}} = \frac{SS_{Tot} - SS_{Res}}{SS_{Tot}}$$ waarbij: $$SS_Y = SS_{Tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$ $$SS_{Mod} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$ $$SS_{Res} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$ **Enkelvoudige lineaire regressie**: populatieniveau $$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-2}$$ **Meervoudige lineaire regressie**: populatieniveau $$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$	**predicties**: fitted(lm) **residuen**: residuals(lm) **SSMod**: sum((fitted(lm)-mean(Y))^2) **SSTot**: var(y)*(n-1)	- R^2 ligt altijd tussen 0 en 1 (het is een proportie) - Hoe hoger R^2, hoe sterker het lineair verband (geeft niet de richting van het verband aan) $R^2 = 1$ → SSMod = SSTot $R^2 = 0$ → SSRes = SSTot

Wanneer welke methode gebruiken?

1 variabele	1) Dichotome variabele: exacte binomiale toets 2) Variabele met meerdere proporties: Pearson chikwadraat toets (homogeniteitstoets) 3) Variabele van minstens intervalniveau: one-sample t-toets
2 variabelen	1) 2 nominale variabelen: Pearson chikwadraat toets (homogeniteitstoets) 2) 2 intervalniveau variabelen: lineaire regressie 3) 1 intervalniveau variabele en 1 dichotome variabele: lineaire regressie of Welch t-toets 4) 1 intervalniveau variabele en 1 nominale variabele: lineaire regressie (of ANOVA) 5) 1 variabele gemeten in 2 populaties: Welch t-toets (indien onafhankelijk) of two-sample paired t-toets