SAMENVATTING · OP TENTAMENNIVEAU

Module 1

Diagnostische analytics en causaliteit

Alles wat je voor de open vragen moet kunnen definiëren, uitleggen, toepassen en beoordelen. Met de modellen als figuur, de vergelijkingen als tabel en de valkuilen gemarkeerd.

WAT ERIN ZIT

01	Centrummaten en spreidingsmaten	2
02	Kansverdelingen en de normale verdeling	2
03	Randomness en de central limit theorem	3
04	Confidence interval	3
05	De empirische cyclus en hypothesen vooraf	3
06	Causal relationship scheme en rollen van variabelen	4
07	Hypothesen formuleren	5
08	Model en variabelen	5
09	Validiteit en onderzoeksdesign	5
10	Voorbeeld uit de School-data	6
11	Valkuilen op een rij	6
→	Zo schrijf je het antwoord	6
→	Checklist: kan ik dit uitleggen?	7
→	Overhoor jezelf	7

TENTAMENKANS

Mediator en moderator ZEKER	94%
Correlatie is geen causaliteit ZEKER	93%
Causaal relatieschema ZEKER	92%
Toetsbare hypothese en nulhypothese formuleren ZEKER	92%
Meetniveaus: nominaal, ordinaal, continu, discreet ZEKER	92%
Hypothesen vooraf formuleren (data dredging) ZEKER	90%
Normale verdeling en kansen berekenen (z-score) ZEKER	90%

Inschatting per onderwerp, op basis van het voorbeeldtentamen en de nadruk in de stof.

KERNBEGRIPPEN

standard deviation · standard error · confidence interval · central limit theorem · causal relationship scheme · mediator · moderator · confounder · reverse causality · internal validity · external validity

OP HET TENTAMEN

Blok 1

10 punten · 180 min totaal

MC	4 meerkeuzevragen	2.5 punten per vraag
7	Blok 7 (25 punten) bestaat uit drie open vragen over R-output.	200 / 250 / 150 woorden

Tentamen: ma 7 dec 2026 · 180 min · Vaktermen blijven Engels

Blok 1 van het tentamen (4 MC, 10 punten) en vraag 7.3 (7 punten, max 150 woorden) komen uit deze module. De rekenregels voor CI en correlatie staan op het formuleblad (zeg dat erbij); de percentages van de normale verdeling (68-95-99,7) staan er niet op, die moet je kennen.

01 Centrummaten en spreidingsmaten

- **Mean** = som / n. **Median** = middelste waarde na ordenen. **Mode** = meest voorkomende waarde.
- **Scheefheid**: symmetrisch geeft mean = median; rechts-scheef geeft mean > median; links-scheef geeft mean < median. Het gemiddelde is gevoelig voor uitschieters, de mediaan niet.
- **Spreiding**: minimum, maximum, range (max – min), MAD (gemiddelde absolute afwijking), variantie, **standard deviation** (wortel uit de variantie), variatiecoëfficiënt = sd / mean, percentielen en IQR (75e minus 25e percentiel).
- **R**: mean(x), median(x), sd(x), var(x), range(x), quantile(x, 0.25), IQR(x), summary(x). In de output van summary() lees je min, 1st Qu., Median, Mean, 3rd Qu., Max.

02 Kansverdelingen en de normale verdeling

- Een **probability mass function** hoort bij discrete variabelen (alle kansen tellen op tot 1); een **probability density function** bij continue (oppervlak onder de curve is 1). De kans op één exacte waarde van een continue variabele is 0, een kans is een oppervlak tussen twee waarden.
- Normale verdeling: symmetrisch, mean = median = mode, bepaald door mu en sigma. 68% binnen 1 sigma, 95% binnen 2 sigma, 99,7% binnen 3 sigma.
- **Z-score**: Z = (x – mu) / sigma. Voorbeeld: wachttijd mu = 7,1 en sigma = 2,3 weken: P(> 11,7) heeft Z = 2, dus ongeveer 2,5%. P(4,8 < wachttijd < 9,4) heeft Z tussen -1 en 1, dus ongeveer 68%.
- **R**: pnorm(x, mu, sigma) geeft de kans links van x (linkerstaart is de standaard), dus 1 – pnorm(11,7; 7,1; 2,3) voor de rechterstaart. qnorm(p) geeft de kritieke waarde, bijvoorbeeld qnorm(0,975) = 1,96.
- Valkuil: Z-score uitrekenen en dan de kans de verkeerde kant op lezen (rechter- versus linkerstaart).

03 Randomness en de central limit theorem

- Een **random variable** is een grootheid waarvan de waarde van toeval afhangt. De kernvraag van de statistiek: is dit **signal** of **noise**?
- Notatie: mu (populatiegemiddelde), sigma (populatie-sd), x-streep (steekproefgemiddelde), s (steekproef-sd), n (steekproefomvang).
- **Central limit theorem**: het gemiddelde van een simple random sample van omvang n is bij benadering normaal verdeeld met verwachting mu en standaardafwijking sigma / wortel(n), ook als de populatie zelf niet normaal is. In de praktijk werkt dit vanaf n van ongeveer 20 tot 30.
- **Standard error** = s / wortel(n) (met sigma als die bekend is). De SE daalt met een grotere n en stijgt met een grotere s. Om de SE te halveren heb je viermaal zoveel waarnemingen nodig.

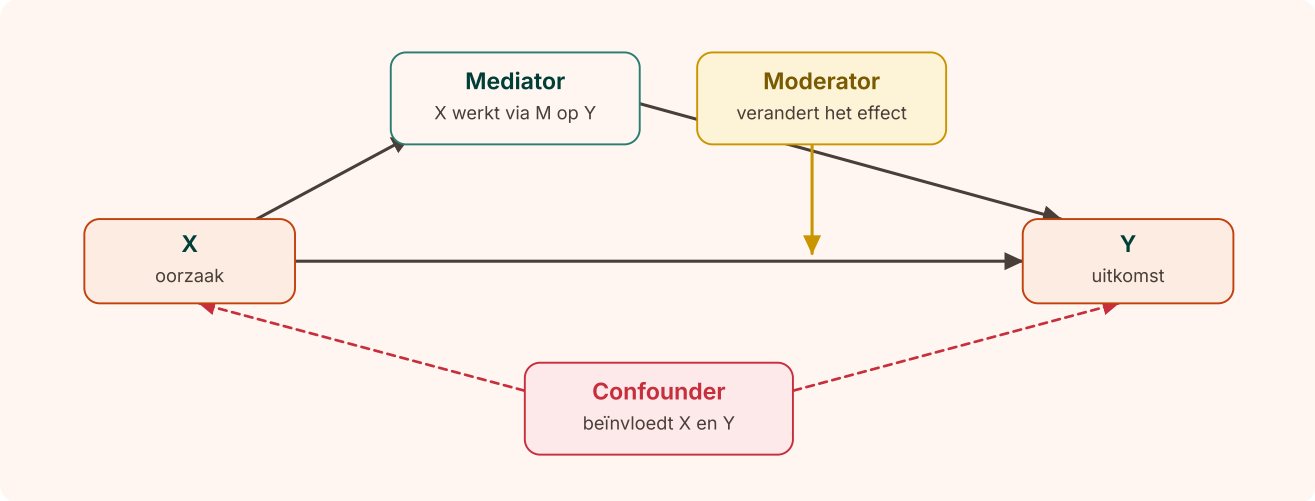
04 Confidence interval

- Formule (op het formuleblad): CI = x-streep +/- t* x s / wortel(n), of z* x sigma / wortel(n) bij bekende sigma of grote n.
- Kritieke waarden: 90% geeft 1,64, 95% geeft 1,96 (vuistregel 2), 99% geeft 2,58. In R: qnorm(1 – alfa/2), bij onbekende sigma qt(0,975, df = n – 1).
- Voorbeeld (Donations, gesimuleerde donaties): x-streep = 36,9, s = 14,3, n = 100, dus SE = 1,43 en CI = [34,1; 39,7].
- **Juiste interpretatie**: bij herhaald steekproeven trekken bevat ongeveer 95% van de intervallen het echte gemiddelde ("we zijn 95% confident"). **Onjuist**: de kans dat mu in dit ene interval zit is 95%; 95% van de data ligt in het interval; 95% kans dat het steekproefgemiddelde van een herhaling erin valt.
- Een groter n of een kleinere s geeft een smaller interval; een hoger betrouwbaarheidsniveau geeft een breder interval.

05 De empirische cyclus en hypothesen vooraf

- Cyclus: onderzoeksvraag, ideeën (theorie, logica, eerdere resultaten), hypothesen, model, data, analyse, conclusie. Drie ingrediënten tussen vraag en data: **causal relationship scheme**, hypothesen, model.
- **Waarom hypothesen vooraf?** (1) Focus en discipline; (2) ze sturen de dataverzameling, zodat je niet de verkeerde data verzamelt of een andere vraag beantwoordt; (3) ze voorkomen **data dredging**: wie maar blijft zoeken, vindt per toeval verbanden; (4) ze beperken confirmation bias achteraf.
- Deductief (toetsend) onderzoek begint met hypothesen, inductief (verkennend) onderzoek zoekt patronen.

06 Causal relationship scheme en rollen van variabelen



FIGUUR 1.1 **Causaal relatieschema.** X beïnvloedt Y direct en via een mediator; een moderator verandert de sterkte; een confounder geeft een schijnverband.

Rol	Betekenis	Voorbeeld
Direct effect	X werkt direct op Y	advertentie, omzet
Mediator	ligt tussen X en Y en verklaart het indirecte effect	advertentie, store visits, omzet
Moderator	bepaalt de sterkte (of het teken) van het effect van X op Y: interactie-effect	seizoen (kerst of niet)
Confounder	beïnvloedt zowel X als Y en kan een schijnverband geven	temperatuur bij ijs en verdrinking
Reverse causality	Y beïnvloedt ook X (simultane effecten)	omzet, budget

- Pijlsymbolen: + verwacht positief, – verwacht negatief, * categorische variabele (moderator als nominale variabele), ? concurrerende hypothesen.
- Herkennen: "verklaart hoe of via welk kanaal" is mediator. "Sterker voor groep A dan voor groep B" is moderator. "Beïnvloedt allebei" is confounder.
- Een mediator zit in het causale pad, een confounder staat erbuiten en veroorzaakt X en Y. Een moderator verandert de sterkte van een pijl, een mediator voegt een extra schakel toe.
- Alleen de relaties waar je naar vraagt komen in het schema. Controlevariabelen zet je er meestal niet in.

07 Hypothesen formuleren

- **Associatief**: "Wie meer uren op social media zit, is minder tevreden over zijn studie." **Causaal**: "Maximaal 30 minuten social media per dag verhoogt de tevredenheid."
- Kies causaal alleen als je design causale inferentie toelaat (experiment: je grijpt in en kijkt wat er verandert). Bij een enquête alleen een associatieve hypothese.
- Een goede hypothese: toetsbaar, met richting, variabelen helder benoemd. De **H0** is de tegenhanger: geen effect (bij een gerichte H1: geen positieve invloed).
- Hypothese per pijl: H1 direct effect, H2 indirect effect via mediator ("positieve indirecte invloed via store visits"), H3 moderator ("groter in de kerstperiode dan in andere seizoenen").

08 Model en variabelen

- Een **model** is een formele, abstracte en subjectieve weergave van de werkelijkheid: geen vaagheid, alleen de essentiële kenmerken, aangenomen relaties. Het is het kader waarbinnen je hypothesen toetst. Een abstract begrip (construct) hoeft niet direct waarneembaar te zijn; je maakt het meetbaar via operationalisatie.
- **Meetniveaus**: nominaal (geen volgorde), ordinaal (volgorde, afstanden betekenisloos), continu en discreet (beide kwantitatief). In R: character of factor, ordered factor, double, integer.
- **Afhankelijke variabele** (Y, uitkomst) en **onafhankelijke variabele** (X, verklarende variabele).

09 Validiteit en onderzoeksdesign

- **Internal validity**: hoe betrouwbaar is het antwoord voor de bronpopulatie zelf? Hangt af van statistiek, steekproefmethode, biases en confounders.
- **External validity**: hoe goed kun je generaliseren naar de target population, andere groepen of andere tijden? Valkuil: een te specifieke bronpopulatie.
- **Experiment** versus **observationele data (enquête)**: een experiment (liefst met loting) geeft sterk bewijs voor causaliteit; observationele data geeft associaties. Dat verklaart waarom causaal geformuleerde conclusies uit een enquête fout zijn.
- Ceteris paribus: causaliteit betekent dat een verandering in X leidt tot een verandering in Y terwijl alle andere invloeden gelijk blijven.

10 Voorbeeld uit de School-data

cor(School$Sat, School$Debts) geeft 0,166 en cor(School$CostTot, School$Debts) geeft 0,424. Dat is samenhang, geen bewijs dat hogere kosten schulden veroorzaken: de verschillen tussen groepen (tapply, aggregate) laten alleen zien dat gemiddelden verschillen.

11 Valkuilen op een rij

- 95% CI lezen als "95% kans dat mu erin zit".
- SE verwarren met sd (SE gaat over het gemiddelde, sd over de data).
- Moderator en mediator door elkaar halen.
- Causale conclusie uit een enquête.
- Hypothese die niet toetsbaar is ("sociale media is belangrijk") of zonder richting.
- Hypothesen pas formuleren nadat je de data hebt bekeken.

VOORBEELDANTWOORD

→ Zo schrijf je het antwoord

Vier bouwstenen, in deze volgorde. Elke steen levert punten op; laat er geen weg.

1. Term
Noem het model of begrip met de Engelse vakterm.

2. Definitie
Geef de definitie of de kenmerken, precies.

3. Casusfeit
Koppel aan een concreet feit uit de casus.

4. Conclusie
Sluit af met een oordeel, advies of gevolg.

OPDRACHT 1 · WOORDLIMIET 75 WOORDEN (WOORDLIMIET AANGENOMEN)

Een onderzoeker ziet dat ijsverkoop en verdrinkingen samen stijgen. Leg uit of ijs verdrinkingen veroorzaakt.

TERM Hier is een confounder in het spel: een variabele die zowel X als Y beïnvloedt. DEFINITIE Een confounder kan een schijnverband geven: X en Y bewegen samen, zonder dat X de oorzaak van Y is. CASUSFEIT Warm weer verhoogt zowel de ijsverkoop als het aantal mensen dat gaat zwemmen. CONCLUSIE Correlatie is geen causaliteit: het verband zegt niets over ijs als oorzaak, tenzij je voor het weer controleert.

65 / 75 woorden

KERNBEGRIPPEN

✓ Checklist: kan ik dit uitleggen?

Vink af wat je uit je hoofd in één zin kunt uitleggen, met de Engelse term.

- ☐ standard deviation
- ☐ standard error
- ☐ confidence interval
- ☐ central limit theorem
- ☐ causal relationship scheme
- ☐ mediator
- ☐ moderator
- ☐ confounder
- ☐ reverse causality
- ☐ internal validity
- ☐ external validity

FLASHCARDS

? Overhoor jezelf

Dek de rechterkolom af en zeg het antwoord hardop.

Vraag	Antwoord
1 Hoe bereken je het gemiddelde (mean)?	Tel alle waarden op en deel door het aantal waarnemingen n.
2 Hoe bereken je de IQR (interkwartielafstand)?	Het 75e percentiel (3rd Qu.) min het 25e percentiel (1st Qu.).
3 Wat geeft pnorm(x, mu, sigma) in R?	Kans links van x (linkerstaart); rechterstaart = 1 – pnorm().
4 Wat is een random variable?	Een grootheid waarvan de uitkomst van toeval afhangt, zoals een steekproefgemiddelde.
5 Hoeveel meer waarnemingen heb je nodig om de standard error te halveren?	Viermaal zoveel, want SE = s / wortel(n) en wortel(4n) = 2 x wortel(n).
6 Hoe lees je een 95% confidence interval juist?	Bij herhaald steekproeven trekken bevat ongeveer 95% van de intervallen het echte gemiddelde.
7 Wat is data dredging?	Zoeken in de data tot je een toevalsverband vindt, zonder hypothesen vooraf.
8 Wat is een confounder?	Derde variabele die X en Y beide beïnvloedt.
9 Wat maakt een hypothese goed?	Toetsbaar, met een richting, en met helder benoemde variabelen.
10 Wat is een nominale variabele?	Categorieën zonder volgorde, zoals land of geslacht (in R: character of factor).
11 Wat is internal validity?	Betrouwbaarheid van het antwoord voor de bronpopulatie zelf.
12 Welke relaties zet je in een causal relationship scheme?	Alleen de relaties uit je onderzoeksvraag; geen controlevariabelen.