SAMENVATTING · OP TENTAMENNIVEAU

Module 2

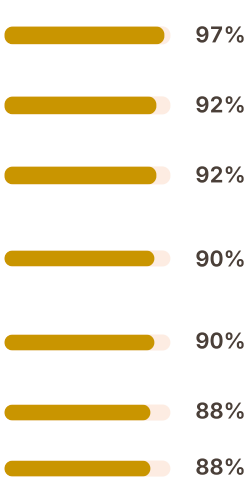
Beschrijvende statistiek, sampling en correlatie

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	Descriptive statistics en pitfalls van shallow statistics	2
02	Visualisatie	2
03	Population, sample en validity	2
04	Sampling methods	3
05	Sampling errors en biases	3
06	Inference: point en interval estimators	4
07	Central limit theorem, standard error en confidence interval	4
08	Covariance	4
09	Correlation	5
10	Correlation is geen causation	5
11	Statistical dependency	5
12	R-output lezen	6
13	Rekenvoorbeeld: van standard error naar interval	6
14	Veelgemaakte fouten	6
→	Zo schrijf je het antwoord	6
→	Checklist: kan ik dit uitleggen?	7
→	Overhoor jezelf	7

TENTAMENKANS



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

KERNBEGRIPPEN

descriptive statistics · population · validity · stratified sampling · selection bias · central limit theorem · confidence interval · covariance · Pearson · Spearman

OP HET TENTAMEN

Blok 2

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

01 Descriptive statistics en pitfalls van shallow statistics



FIGUUR 2.1 **Drie soorten maten.** Beschrijvende statistiek vat een dataset samen.

Descriptive statistics zijn samenvattende maten voor kwantitatieve data: mean, median, standard deviation, interquartile range. Mean groter dan median wijst op een staart naar rechts. Een Z-score is (x – mean) / sd en telt het aantal standaardafwijkingen van het gemiddelde; een waarde met |Z| groter dan 2 valt op. De uitschietergrens van de box plot is Q3 + 1,5 x IQR en Q1 – 1,5 x IQR.

Een samenvatting is nooit neutraal. De retail-voorbeeldkrant maakte drie keuzes: één observatie in plaats van de trend, de vorige maand als referentie in plaats van een jaar eerder, en een as die niet bij nul begint. Vraag je bij elke grafiek af welke keuze erachter zit.

Kwalitatieve data beschrijf je met een frequentietabel, bar chart of pie chart. Een gemiddelde van categoriecodes heeft geen interpretatie.

02 Visualisatie

- Histogram: de verdeling van één kwantitatieve variabele (bins).
- Scatter plot: relatie tussen twee kwantitatieve variabelen, eventueel een derde als kleur.
- Box plot: een kwantitatieve variabele per categorie (median, kwartielen, uitschieters).

03 Population, sample en validity

Je trekt een sample uit een population om daar iets over te zeggen. De drie niveaus zijn target population (waarover je iets wilt zeggen), source population (waaruit je werkelijk trekt) en study sample.

- **Internal validity**: hoe betrouwbaar is het antwoord voor de source population? Hangt af van statistiek, sampling design en biases en confounders.
- **External validity**: in hoeverre generaliseer je naar andere populaties of tijden?

Valkuil in een scriptie: een te specifieke source population kost external validity. **Representativity** betekent dat gemiddelde, spreiding en vorm van de sample lijken op die van de population.

04 Sampling methods



FIGUUR 2.2 **Van populatie naar steekproef.** Een steekproef is alleen bruikbaar als hij representatief is.

- **Simple random sampling**: elk element heeft dezelfde kans. Eenvoudig in theorie, moeilijk in de praktijk (het Literary Digest-voorbeeld van 1936: grote sample, verkeerde selectie).
- **Stratified sampling**: indelen in strata (leeftijd, geslacht, afdeling) en uit elk stratum trekken. Zo is de sample gebalanceerd op die kenmerken.
- **Systematic sampling**: geordende lijst, vaste stapgrootte. Bij N = 2000 en n = 50 neem je elke 40e.
- **Cluster sampling**: clusters (scholen, steden, bedrijven) willekeurig kiezen en daarbinnen data verzamelen.

Het voordeel van de alternatieven: minder moeite bij dataverzameling, vaak een betere schatting.

05 Sampling errors en biases



FIGUUR 2.3 **Van populatie naar steekproef.** Een steekproef is alleen bruikbaar als hij representatief is.

Sampling error is het verschil tussen een sample-statistiek en de populatiewaarde (x-streep versus mu). Non-systematische fouten komen door toeval en hebben geen richting. Systematische fouten zijn biases:

- **Selection bias**: het sampling design over- of onderrepresenteert groepen.
- **Non-response bias**: sommige groepen doen disproportioneel vaak niet mee.
- **Response bias**: antwoorden zijn systematisch niet in lijn met de werkelijkheid. Voorbeeld: dezelfde twee productplaatjes in omgekeerde volgorde gaven 56% voor het eerste plaatje in sample 1 en 58% voor het eerste (andere) plaatje in sample 2.

Een grotere n verkleint de toevalsfout, maar haalt geen bias weg.

06 Inference: point en interval estimators

Inference is een conclusie over de population op basis van een sample. Twee bronnen van onzekerheid: sampling variation en onbekende populatiekenmerken. Een **point estimator** levert één getal; een **interval estimator** levert een interval dat de parameter met een zekere kans bevat.

07 Central limit theorem, standard error en confidence interval



FIGUUR 2.4 **Drie soorten maten.** Beschrijvende statistiek vat een dataset samen.

Voor een random variabele met verwachting mu en eindige variantie sigma kwadraat is het gemiddelde van een simple random sample met grootte n bij benadering normaal verdeeld met gemiddelde mu en standaardafwijking sigma / wortel(n): de **standard error** (SE). In de praktijk werkt dit vanaf n = 20 tot 30.

- 95% confidence interval: x-streep plus of min 1,96 x SE; 90%: 1,64; 99%: 2,58.
- Onbekende sigma: gebruik s en t* uit de t-verdeling: CI = x-streep plus of min t* x s / wortel(n).
- Hoger betrouwbaarheidsniveau betekent een breder interval; vier keer zoveel waarnemingen halveert de SE.
- Interpretatie: de 95% hoort bij de methode, niet bij één interval.

In R: qnorm(0,975) = 1,96; pnorm geeft de linkerstaart, rechterstaart is 1 – pnorm(); qt(0,975, df) geeft t*.

08 Covariance

Cov(X, Y) = som van (x_i – x-streep)(y_i – y-streep), gedeeld door n – 1. Het is de gemiddelde rechthoekoppervlakte rond het zwaartepunt. Positief betekent dat x en y samen stijgen, negatief dat de ene stijgt als de andere daalt. Nadeel: de waarde hangt van de eenheden af en kan alles zijn, dus lastig te interpreteren.

09 Correlation

Pearson: r = Cov(X, Y) / (sd(X) x sd(Y)). Altijd tussen -1 en +1, eenheidsloos. Voorbeeld: Cov = 54, sd's 12 en 9 geeft r = 54 / 108 = 0,5. Wat sterk of zwak is, hangt van de context af.

Drie valkuilen:

1. Een **uitschieter** kan r sterk omhoog of omlaag trekken. Conclusie op basis van r alleen is fout; kijk altijd naar de scatter plot.
2. Een r rond 0 betekent niet dat er geen relatie is; Pearson meet alleen lineaire samenhang (y = x kwadraat geeft r = 0).
3. Bij een niet-lineair maar stijgend verband geeft Pearson een waarde onder 1.

Spearman is de correlatie van de rangen en past bij monotone niet-lineaire verbanden en bij uitschieters.

10 Correlation is geen causation

Een correlatie tussen A en B kan komen door: A veroorzaakt B, B veroorzaakt A (**reverse causality**), toeval (mozzarella en promoties), feed-forward (politie en criminaliteit; fietsen en BMI), of een derde variabele C (**confounder**: ijsverkoop en verdrinkingen via temperatuur).

Een **spurious correlation** is een statistisch verband zonder causaal verband, door toeval of door een confounder. De **partial correlation** meet het verband tussen X en Y na controle voor Z:

r_XY.Z = (r_XY – r_XZ x r_YZ) / wortel((1 – r_XZ kwadraat)(1 – r_YZ kwadraat)).

Een partial correlation die veel kleiner is dan r_XY wijst op een confounder of op een mediator; het design (de causale volgorde in het schema) beslist welke van de twee het is.

11 Statistical dependency

Causaal verband: een verandering in X leidt tot een verandering in Y terwijl alle andere invloeden gelijk blijven (ceteris paribus). Statistisch verband: de gezamenlijke verdeling laat een systematisch patroon zien. Causaliteit is niet hetzelfde als statistische afhankelijkheid; afhankelijkheid kun je objectief met data bepalen, causaliteit is moeilijker.

Twee variabelen zijn statistisch onafhankelijk als de marginale verdeling gelijk is aan de conditionele verdeling: P(koop) = P(koop | blauwe website). Informatie over X heeft dan geen voorspellende waarde voor Y.

12 R-output lezen

Het tentamen vraagt geen R-code, wel het interpreteren van output. In summary() vergelijk je mean en median; in cor.test() lees je r, t, df, de p-waarde en het 95% confidence interval. Bevat het interval 0, dan is de correlatie niet significant bij 5%. Een correlatiematrix lees je per paar; een enkele uitschieter kan een significantie laten verdwijnen of verschijnen.

13 Rekenvoorbeeld: van standard error naar interval

Stel: n = 100, gemiddelde 50 en s = 20 (gesimuleerde data). Dan is SE = 20 / wortel(100) = 2. Het 95% confidence interval is 50 +/- 1,96 × 2, dus [46,08; 53,92]. Wil je de SE halveren naar 1, dan heb je n = 400 nodig: vier keer zoveel, omdat de SE met de wortel van n daalt. Een hoger betrouwbaarheidsniveau (99%: 2,58) maakt het interval breder, een grotere n maakt het smaller.

LET OP

14 Veelgemaakte fouten

- ✖ Zeggen dat een sterke correlatie causaliteit bewijst.
- ✖ Denken dat de 95% een kans voor dit ene interval is.
- ✖ Een grotere sample zien als medicijn tegen selection bias.
- ✖ r = 0 lezen als "geen verband".
- ✖ Een correlatie interpreteren zonder de scatter plot.

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 dataset bevat een paar extreem hoge salarissen. Welke centrummaat past het best en waarom?

TERM De median past het best. DEFINITIE De median is de middelste waarde van de geordende data en wordt niet beïnvloed door uitschieters; de mean wel. CASUSFEIT Een paar extreem hoge salarissen trekken de mean omhoog, zodat die niet meer typisch is voor de meeste werknemers. CONCLUSIE Rapporteer bij een scheve verdeling de median, eventueel naast de mean en de standard deviation.

58 / 75 woorden

KERNBEGRIPPEN

✓ Checklist: kan ik dit uitleggen?

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

- ☐ descriptive statistics
- ☐ population
- ☐ validity
- ☐ stratified sampling
- ☐ selection bias
- ☐ central limit theorem
- ☐ confidence interval
- ☐ covariance
- ☐ Pearson
- ☐ Spearman
- ☐ partial correlation
- ☐ statistical independence

FLASHCARDS

? Overhoor jezelf

Dek de rechterkolom af en zeg het antwoord hardop.

Vraag	Antwoord
1 Waarom is "de verkoop daalde in april" een mogelijk misleidende kop?	Eén maand, een vergelijking met de vorige maand en een gekozen as bepalen het verhaal.
2 Wat betekent een Z-score van 2,5?	De waarde ligt 2,5 standaardafwijkingen boven het gemiddelde.
3 Welk risico loop je met een te specifieke source population?	Je verliest external validity: de uitkomst is moeilijk te generaliseren.
4 Waarom zijn de alternatieven voor random sampling interessant?	Ze verminderen de moeite bij dataverzameling en verbeteren vaak de schatting.
5 Wat ging er mis bij de Literary Digest-peiling van 1936?	Autobezitters, telefoongebruikers en abonnees waren rijker: selection bias, ondanks 2,3 miljoen antwoorden.

Vraag	Antwoord
6 Wat zegt het central limit theorem?	Het sample-gemiddelde is bij benadering normaal verdeeld met gemiddelde mu en sd sigma / wortel(n).
7 Wat is het verschil tussen standaardafwijking en standard error?	Standaardafwijking: spreiding van de data; standard error: spreiding van het sample-gemiddelde.
8 Wanneer gebruik je de t-verdeling in plaats van de normale verdeling?	Als sigma onbekend is en je s gebruikt; df = n – 1.
9 Wat meet de Pearson-correlatie?	Alleen de lineaire samenhang tussen twee kwantitatieve variabelen, tussen -1 en 1.
10 Wanneer gebruik je de Spearman-correlatie?	Bij een stijgend of dalend verband dat niet recht is, of met uitschieters en ordinale data: het werkt op rangen.
11 Wat doet een partial correlation?	Het meet de samenhang tussen X en Y nadat je het effect van een derde variabele Z eruit gehaald hebt.
12 Is statistische afhankelijkheid hetzelfde als causaliteit?	Nee: afhankelijkheid zegt dat variabelen samenhangen; causaliteit vraagt ook om richting en ceteris paribus.