1. Data en hun voorstelling - Beschrijvende statistieken

Oefening 1.1: Gemiddelde januari temperatuur USA

Situatieschets

Voor 56 steden in de Verenigde Staten werd de gemiddelde januari temperatuur T over verscheidene jaren in graden Fahrenheit opgetekend, samen met de breedtegraad B in graden waarop de stad gelegen is. Deze data staan in het bestand gemiddeldejanuaritemperatuur.txt op BB.

T	B	T	B	T	B
0	52	21	38	31	40
1	49	21	38	31	43
8	50	22	42	31	37
9	44	22	44	32	31
9	42	22	49	33	35
11	41	23	39	33	41
11	47	24	41	34	33
12	45	24	40	35	38
13	44	24	35	38	35
13	49	24	39	39	39
14	45	25	35	39	28
14	45	26	42	42	32
15	49	26	37	44	29
18	41	27	40	44	41
20	44	27	38	45	34
20	45	27	40	48	35
21	41	28	37	49	25
21	43	30	40	57	24
				65	23

Vragen

1. Tot welke soort meetschaal behoren deze variabelen?
2. Teken de cumulatieve verdelingsfunctie van de temperatuur. Leid de mediaan af uit deze grafiek.
3. Hoeveel temperaturen zijn kleiner dan 14 °F?
4. We beschouwen de twee kwantitatieve variabelen samen. Maak een scatterplot/puntenwolk van de breedtegraad als functie van de temperatuur.
5. Welk teken denk je dat de correlatie zal hebben? Zal de absolute waarde van de correlatie hoog of laag zijn?
6. Bereken de correlatiecoëfficiënt.

Oefening 1.2: Overlevingsduur proefkonijnen

Situatieschets

Onderstaande tabel toont de overlevingsduur in dagen van 72 proefkonijnen die ten behoeve van een medisch experiment met tuberkelbacillen zijn geïnjecteerd. Deze data staan in het bestand overlevingsduurkonijnen.txt op BB.

43	45	53	56	56	57	58	66	67	73
74	79	80	80	81	81	81	82	83	83
84	88	89	91	91	92	92	97	99	99
100	100	101	102	102	102	103	104	107	108
109	113	114	118	121	123	126	128	137	138
139	144	145	147	156	162	174	178	179	184
191	198	211	214	243	249	329	380	403	511
522	598								

Vragen

1. Maak een geschikte grafiek om de verdeling van de overlevingstijden te beschrijven.
2. Beschrijf de vorm van de verdeling.
3. Geef een numerieke samenvatting van de data bestaande uit de mediaan, het eerste en derde kwartiel, en de grootste en kleinste afzonderlijke waarneming. Bereken het rekenkundig gemiddelde. Geef een maat voor de spreiding uitgedrukt in spreidingsbreedte, interkwartielbreedte, variantie en standaarddeviatie. Bereken de Pearsons scheefheidscoëfficiënt.
4. Zijn er uitschieters?

Oefening 1.3: Verdeling leeftijd

Situatieschets

De verdeling van de leeftijd in de bevolking van een land heeft grote invloed op de sociale en economische omstandigheden. De volgende tabel toont de verdeling van de leeftijd van inwoners in de VS in 1950 en 2075, in miljoenen personen. De gegevens voor 1950 komen van de volkstelling van dat jaar en de data voor 2075 zijn door het Census Bureau gemaakte voorspellingen. Deze data staan in het bestand verdelingleeftijd.txt op BB.

Leeftijdsgroep	**1950**	**2075**
Jonger dan 10 jaar	29.3	34.9
10-19 jaar	21.8	35.7
20-29 jaar	24.0	36.8
30-39 jaar	22.8	38.1
40-49 jaar	19.3	37.8
50-59 jaar	15.5	37.5
60-69 jaar	11.0	34.5
70-79 jaar	5.5	27.2
80-89 jaar	1.6	18.8
90-99 jaar	0.1	7.7
100-109 jaar	-	1.7
Totaal	**150.9**	**310.7**

Vragen

1. Omdat de totale populatie van 2075 veel groter is dan die van 1950, geeft het een duidelijker beeld wanneer men de relatieve frequenties in elke groep vergelijkt dan wanneer men de aantallen vergelijkt. Maak een tabel van de relatieve frequenties van de totale populatie in elke groep voor zowel 1950 als voor 2075.
2. Maak een staafdiagram voor de leeftijdsverdeling van 1950, en beschrijf de algemene eigenschappen van de verdeling. Kijk in het bijzonder naar het aantal kinderen ten opzichte van de rest van de bevolking.
3. Maak een staafdiagram voor de voorspelde leeftijdsverdeling voor het jaar 2075. Wat zijn de belangrijkste veranderingen in de leeftijdsverdeling voor de VS, zoals die voorspeld is voor de periode van 125 jaar tussen 1950 en 2075?

Oefening 1.4: Hoeveelheid calorieën/natrium in hotdogs

Situatieschets

Onderstaande tabel toont de resultaten van laboratoriumonderzoek naar de hoeveelheid calorieën en het aantal milligram natrium in de hotdogs van een aantal belangrijke merken. Er zijn drie typen: uitsluitend rundvlees, vlees (voornamelijk rund- en varkensvlees, maar de overheidsvoorschriften laten een maximum van 15% gevogelte toe) en gevogelte. Deze data staan in het bestand hotdogs.txt op BB.

Rundvlees		**Vlees**		**Gevogelte**	
Calorieën	**Natrium**	**Calorieën**	**Natrium**	**Calorieën**	**Natrium**
186	495	173	458	129	430
181	477	191	506	132	375
176	425	182	473	102	396
149	322	190	545	106	383
184	482	172	496	94	387
190	587	147	360	102	542
158	370	146	387	87	359
139	322	139	386	99	357
175	479	175	507	170	528
148	375	136	393	113	513
152	330	179	405	135	426
111	300	153	372	142	513
141	386	107	144	86	358
153	401	195	511	143	581
190	645	135	405	152	588
157	440	140	428	146	522
131	317	138	339	144	545
149	319				
135	298				
132	253				

Vragen

1. Teken box-plots van de caloriegegevens van de 3 typen hotdogs.
2. Geven de box-plots uitschieters weer? Waarom wel of waarom niet?
3. Welke conclusies kan je trekken uit deze box-plots?
4. Maak een stam- en bladdiagram van de calorische waarden van de 17 merken vleeshotdogs en bespreek de verschillen met de bijbehorende box-plot.
5. Maak een scatterplot/puntenwolk van het aantal calorieën als functie van het natriumgehalte voor de 3 typen hotdogs. Geeft deze een verband tussen beide variabelen aan?

2. Kansrekenen

Oefening 2.1: Werkpremies

Situatieschets

Een bedrijfsleider hecht veel belang aan de werkijver van zijn werknemers. Om die reden werkt hij sedert twee jaar met een premiestelsel. Elk jaar met Nieuwjaar beloont hij de 40 meest vlijtige werknemers uit het totale aantal van 200 werknemers met 1000 euro extra op hun loon van december. Bekeken over de voorbije twee jaar zijn er 13 werknemers die zowel in het eerste als in het tweede jaar de premie gekregen hebben.

Vragen

1. Schrijf de gegevens op in formulevorm.
2. Bereken de kans dat een werknemer minstens 1 keer een premie gekregen heeft.
3. Hoe groot is de kans dat een werknemer enkel in het eerste jaar een premie kreeg?
4. Bereken de kans dat een werknemer die in het eerste jaar reeds een premie kreeg er ook een kreeg in het tweede jaar.
5. Kan je in dit voorbeeld zeggen dat het uitreiken van de premies in de twee verschillende jaren onafhankelijk van elkaar gebeurd is?

Oefening 2.2: Appels

Situatieschets

Een kweker van Jonathan appels wil vier bomen in zijn domein vergelijken. Elk van deze vier bomen werd behandeld met een verschillende meststof en de kweker wil weten welke meststof zwaardere appels geeft dan de andere. Gegeven dat appels respectievelijk aan boom A, B, C en D hangen, is de kans om meer dan 200 gram te wegen respectievelijk gelijk aan 10%, 30%, 65% en 3%. Bovendien kennen we per boom het aantal gewogen appels, namelijk 100 voor A, 200 voor B, 100 voor C en 300 voor D.

Vragen

1. Schrijf de gegevens op in formulevorm.
2. Bereken de kans dat een appel minder dan 200 gram weegt, gegeven dat de appel respectievelijk aan boom A, B, C en D hangt.
3. Bereken de kans dat een willekeurige appel meer weegt dan 200 gram.
4. Als je van een appel weet dat hij meer weegt dan 200 gram, bereken dan de kansen dat dit een appel is afkomstig van boom A, van boom B, van boom C en van boom D.

Oefening 2.3: Autoverzekering

Situatieschets

Bij een verzekeringsmaatschappij zijn 100 autobestuurders verzekerd voor de schade die ze bij een verkeersongeval door eigen schuld aan anderen toebrengen. Onder de verzekerden zijn er 30 die 15% kans hebben om in de loop van het jaar een ongeval te veroorzaken (groep A). Voor de 70 anderen is die kans gelijk aan 5% (groep B). We kiezen willekeurig een verzekerde.

Vragen

1. Schrijf de gegevens op in formulevorm.
2. Bereken de kans dat een verzekerde in de loop van het jaar een ongeval veroorzaakt.
3. Gegeven dat een verzekerde een ongeval heeft veroorzaakt, wat zijn dan de kansen om tot elk van de twee groepen te behoren.
4. Algemene conclusie.

Oefening 2.4: Alcoholtest

Situatieschets

Een automobilist die een aanrijding veroorzaakt moet een bloedproef ondergaan. Nu blijkt dat er in Europa drie verschillende bloedproeven bestaan, elk van mekaar verschillend door gebruik te maken van andere chemische producten. De Europese minister die bevoegd is voor veiligheid in het verkeer wil deze bloedproeven vergelijken. Uit studies blijkt dat 1% van de automobilisten die een aanrijding veroorzaken onder invloed is. De drie verschillende bloedproeven hebben respectievelijk 75%, 85% en 65% kans dat het resultaat positief is als de automobilist onder invloed is. De bloedproeven geven respectievelijk met 5%, 10% en 2% kans aan dat het resultaat positief is als de automobilist niet onder invloed is.

Vragen

1. Schrijf de gegevens op in formulevorm.
2. Welke voorwaardelijke kansen kan je berekenen om de kwaliteit van de bloedproeven na te gaan? Bereken deze kansen.
3. Welke bloedproef is de beste? Wat is de algemene indruk over de kwaliteit van de drie bloedproeven? Kan je een oplossing hiervoor bedenken?

Oefening 2.5: Serie- en parallelschakeling

Situatieschets

(a) Beschouw een schakeling die uit vier componenten bestaat die serieel geplaatst zijn. Stel dat elk van de vier componenten een faalkans heeft van 1%, waarbij het al dan niet falen van een component onafhankelijk wordt verondersteld van het al dan niet falen van de andere componenten.

(b) Beschouw een schakeling die uit drie componenten bestaat die parallel geplaatst zijn. Stel dat elk van de drie componenten een faalkans heeft van 5%, waarbij het al dan niet

falen van een component onafhankelijk wordt verondersteld van het al dan niet falen van de andere component.

Vragen

1. Bereken de kans dat het systeem uit (a) werkt.
2. Bereken de kans dat het systeem uit (b) werkt.

3. Univariate kansvariabelen - Kengetallen

Oefening 3.1: Pruimentaarten

Situatieschets

Uit ervaring weet een bakker dat de dagelijkse vraag (V) naar pruimentaarten de volgende kansverdeling heeft:

v	0	1	2	3	4	5	6	7
P(V=v)	0.08	0.10	0.10	0.14	0.17	0.17	0.14	0.10

Het produceren van een pruimentaart kost de bakker 2 euro. De verkoopprijs bedraagt 3 euro en taarten die op het einde van de dag niet verkocht zijn hebben nog een restwaarde van 1 euro bij verwerking in andere bereidingen

Vragen

1. Bereken de verwachte waarde van de dagelijkse vraag naar pruimentaart.
2. Noem G het aantal geproduceerde pruimentaarten. Geef de formules voor de dagelijkse winst W.
3. Maak een tabel waarin voor elke waarde van V en voor elke waarde van G de verkregen winst staat. Voeg bovendien een kolom toe waarin je voor elke waarde van G de te verwachten winst zet. Voeg ten slotte ook 3 rijen toe waarin je voor elke waarde van V de getalwaarde geeft van P(V=v), van P(V≥v) en van de cumulatieve verdelingsfunctie $F_V(v)$.
4. Teken de cumulatieve verdelingsfunctie van V.
5. Hoeveel pruimentaarten moet de bakker dagelijks produceren opdat de verwachte dagelijkse winst maximaal is?

Oefening 3.2: Metaalzaagmachines

Situatieschets

Het management van een aluminium-producerend bedrijf overweegt de aankoop van een nieuwe metaalzaagmachine. Deze machine zou dagelijks 26 uren moeten draaien, en aangezien dit niet mogelijk is, is men genoodzaakt om twee machines te kopen, één die dagelijks 20 uren draait en één die dagelijks 6 uren draait. Het dagelijks aantal herstellingen aan deze zaagmachine kan beschouwd worden als een kansvariabele X met verwachte waarde 0.1T en met variantie 0.2T. Hierbij staat T voor het aantal uren per dag dat deze machine operationeel zal zijn. Het operationeel houden ervan geeft een kost C(T) gegeven door: $C(T)=10T+30X^2$

Vragen

1. Schrijf de gegevens op in formulevorm en bereken de te verwachten kost eveneens in formulevorm (in functie van T).
2. Bereken de te verwachten kost voor het operationeel houden van de machines in de geschetste situatie, d.i. één zaagmachine die dagelijks 20 uren draait en één die dagelijks 6 uren draait.

3. Maak een grafiek van de te verwachten kost als functie van T, waarbij T discrete waarden aanneemt van 0 tot en met 24 uren met stappen van 1 uur.
4. Op welke manier kan het bedrijf haar onderhoudskost verkleinen?

4. Belangrijke discrete kansverdelingen

Oefening 4.1: Assemblagelijn

Situatieschets

Van een assemblagelijn is geweten dat ze 12% defecte producten aflevert.

Vragen

1. Een lukraak gekozen afgewerkt product wordt aan een inspectie onderworpen. Wat is de kans dat het om een defect product gaat?
2. Men kiest lukraak en onafhankelijk van elkaar twee afgewerkte producten. Wat is de kans dat het allebei defecte producten zijn?
3. Men kiest lukraak en onafhankelijk van elkaar 20 afgewerkte producten. Wat is de kans dat 5 producten defect zijn?

Oefening 4.2: Basketbalspeler

Situatieschets

Rick is een professionele basketbalspeler. Uit zijn prestaties in het verleden leert men dat hij bij elke vrije worp 75% kans op succes heeft. In een cruciale wedstrijd gooit Rick 12 vrije worpen en mist daarvan 5. De fans denken dat hij miste omdat hij nerveus was. Is het voor Rick ongebruikelijk om zo slecht te presteren?

Vragen

1. Bereken de kans op het missen van ten minste 5 vrije worpen wanneer er 12 gegooid worden.
2. Vergelijk de prestatie van Rick met zijn te verwachten aantal gemiste vrije worpen op 12.

Oefening 4.3: Overboekingen

De uitbater van een hotel verhuurt 200 kamers. De voorbije jaren noteerde de manager dat 5% van alle mensen die een kamer reserveerden uiteindelijk niet opdagen. Daarom begint hij systematisch meer boekingen toe te laten dan er kamers beschikbaar zijn. Hij wil evenwel niet onbeperkt overboekingen verrichten uit schrik potentiële klanten te moeten ontgoochelen. Daarom wil de manager niet meer dan 1% risico lopen om één of meerdere gasten te moeten weigeren. Hoeveel overboekingen mag de manager dan toelaten?

Oefening 4.4: Typfouten

Veronderstel dat er in een 500 pagina's dik manuscript 200 typfouten random verdeeld zijn. Wat is dan de kans dat er op een gegeven pagina exact 3 typfouten staan? Neem aan dat het aantal typfouten per pagina Poisson verdeeld is.

Oefening 4.5: Bloemenverkoper

Een bloemenverkoper verkoopt gemiddeld 20 bloemen per avond. Omdat de bloemen na een avond rondzeulen ’s anderendaags waardeloos zijn, wil de verkoper de voorraad die hij dagelijks inkoopt minimaal houden. Om zijn winst te maximaliseren wil hij evenwel niet al te vaak uitverkocht raken. Veronderstel dat de gevraagde hoeveelheid bloemen op één avond Poisson verdeeld is.

Vragen

1. Maak een grafiek van de kansverdeling.
2. Hoe groot moet de voorraad zijn om op 90% van de avonden aan de vraag te voldoen?

5. Belangrijke continue kansdichtheden

Oefening 5.1: Muizen

Als men muizen met een bepaald product behandelt, vindt men dat de concentratie aan bilirubine in het bloed normaal verdeeld is met gemiddelde 7.5 en onbekende standaardafwijking. Uit proeven blijkt dat er 80.3% kans is dat de concentratie zal begrepen zijn tussen 6 en 9. Bepaal de standaardafwijking

Oefening 5.2: Melkfabriek

In een melkfabriek worden flessen van 1 l machinaal gevuld. Over een voldoende lange periode bekeken is het gemiddelde dat gevuld wordt 1 l met een standaardafwijking van 1 cl. Onderstel dat de vullingsgraad normaal verdeeld is.

Vragen

1. Zoek de waarschijnlijkheid dat er meer dan 1.01 l in een fles zit.
2. Wat moet de gemiddelde vullingsgraad zijn opdat de kans dat er meer dan 1.01 l in een fles zit slechts 5% is?

Oefening 5.3: Reistijd

Een student die om 7u30 thuis vertrekt, is in 2.5% van de gevallen te laat voor de les die om 8u30 begint. Indien hij vertrekt om 7u25 is hij slechts in 1% van de gevallen te laat. Aannemend dat de reistijd normaal verdeeld is, om hoe laat moet hij dan ten laatste thuis vertrekken om in niet meer dan 0.5% van de gevallen te laat te komen?

6. Multivariate kansvariabelen - Covariantie, correlatie en variantie van lineaire functies

Oefening 6.1

Situatieschets

Stel dat de gezamenlijke kansverdeling voor X='Inkomen' en Y='Opleidingsniveau' gegeven wordt door onderstaande tabel.

		Inkomen			
		1	2	3	4
Opleidingsniveau	1	0.23	0.40	0.1	0.02
	2	0.01	0.09	0.1	0.05

Vragen

5. Bepaal de marginale kansverdeling voor X en Y.
6. Zijn de kansvariabelen X en Y onafhankelijk?
7. Leid de voorwaardelijke kansverdeling $p_{X|Y}(x|y=1)$ af.
8. Bepaal de variantie van X en Y, de verwachte waarde van Z=XY en de covariantie tussen X en Y.
9. Definieer op basis van X en Y een nieuwe kansvariabele Z=X+2Y. Bereken de verwachtingswaarde van Z. Bereken de variantie van Z.
10. Bereken de verwachtingswaarde van $Z=X+2Y^2$.

7. Schatten van populatieparameters

Oefening 7.1

Situatieschets

In de file 'pollutiedataset.txt' die jullie terug vinden op BB worden steekproefgegevens weergegeven van 60 Amerikaanse steden. In deze studie werden waarnemingen omtrent luchtvervuilingspotentieel, socio-economische factoren en klimaatsfactoren gemeten. In deze vraag concentreren we ons op de variabele 'pop' (X) die het aantal inwoners weergeeft.

Vragen

1. Maak een histogram voor de variabele X. Verwacht je op basis van dit histogram dat X normaal verdeeld is?
2. Maak een schatting voor het populatiegemiddelde en de populatievariantie.
3. Welke kansdichtheid verwacht je voor het steekproefgemiddelde? Geef een schatting voor het gemiddelde en de variantie.
4. Bepaal de kans dat de gemiddelde bevolkingsgrootte over 59 Amerikaanse steden groter is dan 2 miljoen inwoners.
5. Bepaal de kans dat de totale bevolking in een steekproef van 59 steden de waarde 59000000 overschrijdt.

Oefening 7.2

Het aanvaardbaar gehalte van een zeker bestanddeel in de lucht is bepaald op 7.7. Het ware gehalte X is normaal verdeeld met gemiddelde waarde 7.6 en variantie 0.0016. De metingen van dit gehalte bevatten een fout Y met gemiddelde waarde 0 en variantie 0.0009. Deze fout is normaal verdeeld en onafhankelijk van het gehalte.

1. Bepaal de waarschijnlijkheid dat 1 enkele meting een waarde oplevert die groter is dan 7.7.
2. Bepaal de waarschijnlijkheid dat het rekenkundig gemiddelde van drie verschillende metingen op drie verschillende tijdstippen groter is dan 7.7.
3. Bepaal de waarschijnlijkheid dat het rekenkundig gemiddelde van drie metingen op een zelfde tijdstip en op een zelfde plaats groter is dan 7.7.

Oefening 7.3

Bij een schattingsprobleem voor de proportie π beschikken we over:
- Steekproefwaarnemingen X die binomiaal verdeeld zijn met parameters n=100 en π/2
- Steekproefwaarnemingen Y die binomiaal verdeeld zijn met parameters n=150 en 2π/3

Als schatters voor de proportie π beschouwen we de schatters $\hat{\pi}_1 = X/50$ en $\hat{\pi}_2 = Y/100$.

a) Bewijs dat beide schatters zuivere schatters zijn voor de proportie.

b) Bepaal de relatieve efficiëntie van $\hat{\pi}_2$ ten opzichte van $\hat{\pi}_1$.

8. Intervalschatters

Oefening 8.1: Eiken

De USA kent 50 soorten eiken, waarvan 28 soorten aan de Atlantische kust voorkomen en 11 in de omgeving van California. Voor elke soort werd het volume van de eikels gemeten. Er werd vermoed dat dit volume een effect heeft op de grootte van het geografische gebied waarop de bomen vookomen. Het lijkt aannemelijk dat de grootte van de zaden van planten een effect heeft op dat geografisch gebied omdat grotere eikels weggedragen worden door grotere dieren, die in een groter territoriaal gebied leven. De gegevens werden gepubliceerd in *Journal of Biogeography (1990), volume 17, 327-332*. De gegevens zijn beschikbaar op BB in de file 'eiken.txt'. Er werden metingen gedaan voor de 39 eikensoorten in de Atlantische en Californische regio. Vier variabelen werden voor elke soort bijgehouden: Atlantische of Californische regio, de grootte van het geografische gebied waar de soort voorkomt (in 100 km^2), het volume van de eikel (in cm^3) en de hoogte van de boom (in m).

Beschouw de dataset 'eiken.txt'. Voor de Californische eiken veronderstellen we dat de grootte van de gebieden, X, normaal verdeeld is.

Vragen

7. Leid een 95% betrouwbaarheidsinterval af voor het populatiegemiddelde.
8. Leid een 90% betrouwbaarheidsinterval af voor het populatiegemiddelde.
9. Leid een 95% betrouwbaarheidsinterval af voor de populatievariantie.
10. Leid een 90% betrouwbaarheidsinterval af voor de populatievariantie.

Oefening 8.2

Voor welke waarde van a is de gemiddelde gekwadrateerde afwijking (GGA) van volgende schatter van μ het kleinst?

$$\hat{\mu} = aX_1 + (1-a)X_2$$

met X_1 en X_2 onafhankelijke steekproefwaarnemingen uit een populatie met populatiegemiddelde μ en populatievariantie σ^2.

Oefening 8.3

Beschouw de dataset 'pollutiedataset.txt' In deze vraag concentreren we ons op de variabele 'Mortality' (X). Op basis van statistische testen mogen we er van uitgaan dat X normaal verdeeld is. Stel dat men uit historisch onderzoek bovendien weet dat σ^2 gelijk is aan 3845.

Vragen

1. Leid een 95% betrouwbaarheidsinterval af voor het populatiegemiddelde.
2. Hoe breed is dit betrouwbaarheidsinterval?
3. Hoe groot zou de steekproef moeten zijn opdat de breedte de waarde 20 niet overschrijdt?
4. Leid een 95% betrouwbaarheidsinterval af voor het populatiegemiddelde wanneer de populatievariantie niet gekend is.

9. Hypothesetoetsen voor één populatie

Oefening 9.1

Beschouw de dataset 'pollutiedataset.txt'. Ga voor het gemiddelde μ van de variabele 'Mortality' na welk van de hypothesen

$$\mu = 950$$
$$\mu < 950$$

de voorkeur krijgt (significantieniveau α=0.05). Maak hiervoor gebruik van drie benaderingen (kritieke waarden, toetsingsgrootheid, p-waarde). Vergelijk de resultaten met deze verkregen op basis van het commando t.test in R. De gegevens van de variabele 'Mortality' zijn normaal verdeeld (zie voorgaande wb).

Oefening 9.2

Wat is de waarschijnlijkheid dat de variantie van een steekproef (n=17) zou begrepen zijn tussen de helft en het dubbele van de populatievariantie? De steekproef is genomen van een normaal verdeelde populatie.

Oefening 9.3

We gooien een muntstuk 50 keer en verkrijgen 17 keer kop. Is het verantwoord te vermoeden dat het muntstuk vervalst is? Beschouw α = 1% en α = 5%.
Formuleer de te toetsen hypothesen, geef de waarde van en bereken de toetsingsgrootheid en een uitdrukking voor de p-waarde

Oefening 9.4

Een onderzoeker heeft over een lange periode gevonden dat zijn drukmeter een standaarddeviatie heeft van σ_0 = 1 Pa. In een recent experiment noteerde hij de volgende resultaten: 38.5 Pa, 41.5 Pa, 40.0 Pa, 42.2 Pa en 37.8 Pa. Is de precisie van de drukmeter veranderd? Gebruik α = 5 %. De gegevens zijn normaal verdeeld. Formuleer de te testen hypothesen. Geef een uitdrukking voor en bereken de toetsingsgrootheid en de p-waarde.

Oefening 9.5

Beschouw de dataset 'pollutiedataset.txt'. Ga na of de gegevens 'Mortality' en 'NOxPot' (stikstofoxidepotentieel) mogelijk uit een normaal verdeelde populatie afkomstig zijn gebruik makend van een kwantieldiagram en een Shapiro wilk test.

Oefening 9.6

Ga voor de mediaan Me van de variabele 'NOxPot' na welk van de hypothesen

$$Me = 15$$
$$Me < 15$$

de voorkeur krijgt (significantieniveau α=0.05).

Oefening 9.7

Men onderzoekt het aantal jongens in 80 families met 5 kinderen. De waargenomen distributie is in onderstaande tabel weergegeven. Toets de hypothese dat de gegevens afkomstig zijn uit uit een binomiale distributie met parameter π = 0.4. Neem α = 0.05.

Aantal jongens in een gezin X	0	1	2	3	4	5
Aantal gezinnen met X jongens	13	18	20	18	6	5

10. Hypothesetoetsen voor twee populaties

Oefening 10.1

Beschouw de dataset 'pollutiedataset.txt'. Ga na of de mortaliteit gemiddeld gezien hoger is voor steden met een lager opleidingsniveau ('Education'<11) dan voor steden met een hogere graad van opleiding ('Education'>=11).

Formuleer de hypothesen. Maak gebruik van R commando's voor het toetsen van hypothesen. Bespreek de output.

Vergeet niet om de voorwaarden bij het toepassen van een toets na te gaan (normaliteit en gelijkheid van varianties wanneer 1 van de steekproeven kleiner is dan 30).

Oefening 10.2

Melganzenvoet is een veelvoorkomend onkruid in korenvelden. Een onderzoeker zaaide op identieke wijze graan in 8 veldjes, vervolgens heeft hij de graanrijen handmatig gewied om ervoor te zorgen dat in vier willekeurig gekozen veldjes geen onkruid groeit en in de andere 4 veldjes precies 9 melganzenvoeten per meter rij. Hieronder volgt de graanopbrengst (bushels per acre) van elk van de veldjes:

Onkruid per meter	**Opbrengst (bushels/acre)**			
0	166.7	172.2	165.0	176.9
9	162.8	142.4	162.7	162.4

Veroorzaken kleine hoeveelheden onkruid een lagere graanopbrengst?
De steekproeven zijn te klein om normaliteit op een juiste manier vast te stellen.

Oefening 10.3

Een bandenfabrikant wenst het gedrag van twee verschillende types autobanden, A en B, na te gaan. Daartoe beslist hij op willekeurige wijze 1 band van type A en 1 band van type B te monteren op de achterwielen van 5 auto's. Daarna wordt met elke wagen een testrit gemaakt van een voorafbepaald aantal kilometer en voor elke band wordt de slijtage gemeten. De resultaten zijn:

Auto	Band A	Band B
1	10.6	10.2
2	9.8	9.4
3	12.3	11.8
4	9.7	9.1
5	8.8	8.3

Kan op basis van deze meetresultaten met significantieniveau 0.05 besloten worden dat er geen verschil in slijtage bestaat tussen de twee types van banden? De data staan in banden.txt.

Oefening 10.4

Twee automerken, Mercedes en Volvo, worden getest om te zien welke van beide de beste resulaten haalt op een bepaald soort crash. De grootte van de schade wordt vervolgens bepaald volgens een schaal van 0 tot 100 (100 is de grootste schade). Zulke testen zijn uiteraard vrij duur en daarom wordt het aantal testen beperkt (5 Mercedes en 4 Volvo). X en Y meten resp. de schade aan de Mercedes en aan de Volvo (X en Y mogen normaal verdeeld verondersteld worden met dezelfde variantie). Veronderstel de volgende resultaten: $\bar{x}$ = 43.3 en s_x^2 = 4.5 voor Mercedes en $\bar{y}$ = 45.2 en s_y^2 = 3.8 voor Volvo. Mercedes beweert dat zijn wagens gemiddeld beter presteren (d.i. minder schade oplopen) op de test dan Volvo. Toets deze bewering op het 5% significantieniveau.

Oefening 10.5

In een test met 2 verschillende soorten plastiek bekwam men volgende resultaten voor de standaardafwijking: s_A = 300 (n_a=11) en s_B = 200 (n_b=21). Kan men zeggen dat de standaardafwijking van plastiek A groter is dan deze van plastiek B aan het α = 5% significantieniveau. Je mag veronderstellen dat de variabelen A en B normaal verdeeld zijn.

Oefening 10.6

In de file 'vitamineC.txt' staan gegevens over het gehalte aan vitamine C in 27 zakken met een mengsel van tarwe en soja. De observatievector 'Fabriek' bevat de gegevens zoals gemeten in een fabriek. De observatievector 'Haiti' bevat de gegevens van de overeenkomstige zak vijf maanden later na transport naar Haïti. We willen weten of er tijdens transport en opslag een verlies aan vitamine C is opgetreden. Maak zowel gebruik van een T-toets en een niet-parametrische toets.

Oefening 10.7

Een schoenfabrikant wenst na te gaan of de kwaliteit van een nieuw en goedkoper materiaal voor schoenzolen (B) minder is in kwaliteit dan een oud en duurder materiaal (A). De fabrikant wil de proef opzetten met 10 kinderen. Hij kan de proef op verschillende manieren opzetten:

- 5 kinderen krijgen schoenen uit materiaal A en 5 kinderen krijgen schoenen uit materiaal B
- Ieder van de 10 kinderen krijgt een schoen in materiaal A en een schoen in materiaal B

Toets de nulhypothese voor de twee proefopzetten met de volgende meetgegevens voor de slijtage:

A (oud)	13.2	8.2	10.9	14.3	10.7	6.6	9.5	10.8	8.0	14.1
B (nieuw)	14	8.8	11.2	14.2	11.8	6.4	9.8	11.3	9.3	13.6

11. Hypothesetoets voor meer dan twee populatiegemiddeldes

Oefening 11.1

Beschouw een veldproef met als doel het vergelijken van de opbrengsten (uitgedrukt in kg/75 m^2) aan vlasvezels voor 7 verschillende rassen, 4 herhalingen per ras. De waarnemingen staan op BB in de file 'rassen.txt'.

1. Zijn er verschillen in de gemiddelde opbrengsten van de 7 rassen? Formuleer de hypothesen.
2. Indien er verschillen zijn, welke rassen verschillen van elkaar?
3. Maak een boxplot om eventuele verschillen te visualiseren.
4. Ga de voorwaarden voor het toepassen van ANOVA na (normaliteit van de residuele waarden + gelijkheid van varianties).

Oefening 11.2

De aanwezigheid van schadelijke insecten in akkers wordt ontdekt door met een kleverige substantie besmeerde borden te plaatsen en de op de borden gevangen insecten te onderzoeken. Om na te gaan welke kleuren het meest aantrekkelijk zijn voor graanbladhaantjes, plaatsen onderzoekers in juli voor elk van vier kleuren zes borden in een haverveld. De tabel hieronder geeft data over het aantal opgevangen bladhaantjes. De waarnemingen staan op BB in de file 'insecten.txt'.

		Opgevangen insecten					
Kleur	**Citroengeel**	45	59	48	46	38	47
	Wit	21	12	14	17	13	17
	Groen	37	32	15	25	39	41
	Blauw	16	11	20	21	14	7

1. Bereken de gemiddelden en de varianties voor de vier kleuren.
2. Formuleer H_0 en H_a voor een ANOVA op deze gegevens en verklaar in woorden wat ANOVA in deze situatie toetst.
3. Gebruik R om de ANOVA uit te voeren. Bepaal de toetsingsgrootheid en de overschrijdingskans. Wat is uw conclusie?
4. Ga de voorwaarden voor het toepassen van ANOVA na.
5. Indien er verschillen zijn tussen verschillende kleuren, ga dan na welke kleurparen significant verschillend zijn. Welke kleur zou u voor het aantrekken van bladhaantjes aanbevelen?

Oefening 11.3

Hoe beïnvloeden nematoden (microscopische wormen, aaltjes) de groei van planten? Een botanicus prepareert 16 identieke potten en plaatst verschillende aantallen nematoden in de potten. In elke pot doet hij een tomatenplantje. Hieronder staan de gegevens over de toename van de lengte van de plantjes (in centimeters), 16 dagen na het planten. Deze zijn ook terug te vinden in de file 'nematoden.txt'.

		Groei van de plantjes			
Nematoden	**0**	10.8	9.1	13.5	9.2
	1000	11.1	11.1	8.2	11.3
	5000	5.4	4.6	7.4	5.0
	10000	5.8	5.3	3.2	7.5

1. Zijn er significante verschillen? Indien er verschillen zijn, ga dan na welke behandelingen significant verschillend zijn.
2. Veronderstel dat de eerste waarneming bij vergissing als 108 werd ingevoerd in plaats van 10.8. Vergelijk deze uitkomsten met de resultaten die u met de juiste gegevensverzameling hebt gevonden. Wat verduidelijkt dit over de invloed van uitschieters in een ANOVA?
3. De logaritme-transformatie wordt vaak toegepast op variabelen als plantengroei. In veel gevallen heeft dit tot gevolg dat de standaardafwijkingen van groepen meer op elkaar gelijken en de gegevens in groepen normaler lijken. Voer de ANOVA opnieuw uit met de logaritmen van de juiste waarden (commando log10 in R). Vergelijk deze resultaten met de resultaten die u bij het analyseren van de onbewerkte gegevens verkregen hebt.

Oefening 11.4

Een bioloog wil nagaan of er variatie is in het suikergehalte van appels tussen verschillende rassen. Hiervoor bepaalt hij het suikergehalte in 3 appels voor 5 random geselecteerde rassen. Zijn er significante verschillen? De waarnemingen staan op BB in de file 'suikergehalteappels.txt'.

		Suikergehalte		
		1	**2**	**3**
Ras	**1**	32.2	34.1	35.0
	2	28.4	31.5	30.6
	3	19.3	26.1	24.7
	4	26.3	27.1	26.5
	5	48.1	52.7	54.1

12. Lineaire regressie

Oefening 12.1

Beschouw de dataset 'pollutiedataset_lr.txt'.

1. Ga door middel van een scatterplot na dat het verband tussen 'Mortality' (mortaliteit) en 'S02Pot' (zwaveldioxidepotentieel) niet lineair is.
2. Ga door middel van een scatterplot na dat dit verband rechtlijniger wordt na een logaritmische transformatie van 'S02Pot'.
3. Geef het verband tusssen 'Mortality' en 'S02Pot'.
4. Maak een scatterplot met bijbehorende regressierechte.
5. Construeer 95% betrouwbaarheidsintervallen voor de geschatte regressiecoëfficiënten.
6. Ga door middel van een hypothesetoets na of het lineaire verband tussen 'Mortality' en 'S02Pot' significant is.
7. Maak een schatting van de variantie van de residu's.
8. Voorspel de verwachte mortaliteit wanneer log(S02Pot) gelijk is aan log(250).
9. Construeer een 95% betrouwbaarheidsinterval voor het gemiddeld aantal doden bij log(S02Pot) = log(250). Wat is de betekenis hiervan?
10. Construeer een 95% predictie-interval voor het aantal doden bij log(S02Pot) = log(250). Wat is de betekenis?
11. Bereken en visualiseer het 95% betrouwbaarheids- en predictie-interval voor alle waarden van log(S02Pot) tussen 0 en 6.
12. Voor welke waarde log(S02Pot) is het betrouwbaarheidsinterval het smalst?
13. Ga de modelonderstellingen voor het toepassen van lineaire regressie na.

Oefening 12.2

De Toren van Pisa is een architectonisch wonder. Ingenieurs die zich zorgen maken over de stabiliteit van de toren hebben uitvoerig onderzoek verricht naar het toenemende hellen van de toren. Metingen van de helling in de loop der tijd leveren veel nuttige informatie. De onderstaande tabel geeft metingen voor de jaren 1975 tot 1987. De variabele 'helling' geeft het verschil tussen het punt waar de toren zou zijn als hij recht stond, en het punt waar hij zich feitelijk bevindt. De gegevens zijn gecodeerd in tienden van millimeters meer dan 2.9 meter, zodat de helling van 1975, die gelijk was aan 2.9642 meter, in de tabel verschijnt als 642. Onderstaande data staan in torenpisa.txt.

Jaar	75	76	77	78	79	80	81	82	83	84	85	86	87
Helling	642	644	656	667	673	688	696	698	713	717	725	742	757

1. Ga door middel van een scatterplot na of het verband tussen 'Helling' en 'Jaar' lineair lijkt.
2. Hoe luidt de vergelijking van de regressierechte?
3. Geef een 95% betrouwbaarheidsinterval voor de verwachte veranderingssnelheid (in tienden mm per jaar) van de helling.
4. In 1918 was de helling 2.9060 meter (de gecodeerde waarde is 60). Gebruik de regressierechte van de jaren 1975 tot 1987 om de voorspelde waarde voor het jaar 1918 te bepalen. Merk op dat u de gecodeerde waarde 18 voor dat jaar moet gebruiken.

5. Is deze voorspelde waarde in overeenstemming met de gemeten waarde in het jaar 1918?
6. Ga de modelonderstellingen voor het toepassen van lineaire regressie na.

Oefening 12.3

Het menselijk lichaam verbruikt bij inspanning meer zuurstof dan in rusttoestand. Om de spieren van zuurstof te voorzien, moet het hart sneller kloppen. De hartslag is gemakkelijk te meten, maar het meten van de hoeveelheid opgenomen zuurstof vereist ingewikkelde apparatuur. Als de zuurstofopname (VO2) nauwkeurig kan worden voorspeld uit de hartslag (HR) kunnen bij onderzoek de voorspelde waarden de feitelijk gemeten waarden vervangen. Helaas zijn niet alle menselijke lichamen gelijk, daarom is er niet 1 enkele voorspellingsvergelijking die voor alle mensen geldig is. Onderzoekers kunnen echter voor 1 persoon zowel HR als VO2 meten bij variërende inspanningsniveaus, en een regressierechte berekenen waaruit voor die persoon de zuurstofopname kan worden voorspeld uit zijn hartslag. Hier volgen de gegevens voor die ene persoon. Deze zijn ook terug te vinden in het bestand zuurstof.txt.

HR	94	96	95	95	94	95	94	104	104	106
VO2	0.673	0.753	0.929	0.939	0.832	0.983	1.049	1.178	1.176	1.292

HR	108	110	113	113	118	115	121	127	131
VO2	1.403	1.499	1.529	1.599	1.749	1.746	1.897	2.040	2.231

1. Hoe luidt de vergelijking van de regressierechte?
2. Teken de gegevens samen met de regressierechte.
3. Toets de nulhypothese dat de helling van de rechte gelijk is aan 0. Verklaar in woorden de betekenis van uw conclusie uit deze toets.
4. Bereken het 95% predictie-interval voor de zuurstofopname van deze persoon, voor een toekomstige gebeurtenis waarbij zijn hartslag 95 bedraagt.
5. Ga de modelonderstellingen voor het toepassen van lineaire regressie na.