Data Science

# Sessie 1 – Voorstellen van gegevens & centrummaten

## Soorten statistiek

### Beschrijvende statistiek

* Gebruikt numerieke en grafische methoden om patronen in een gegevensverzameling te ontdekken, samen te vatten en op een overzichtelijke manier te presenteren.

### Verklarende statistiek

* Gebruikt steekproefgegevens voor het schatten, nemen van beslissingen en voorspellen.
* = inductieve statistiek = inferentiële statistiek

### Experimentele eenheid

* Een object waarvan we gegevens vastleggen

### Populatie

* Verzameling van eenheden die we willen bestuderen

## Statistische methoden als analyse-instrument

### Variabele

* Kenmerk of eigenschap van een eenheid uit een populatie

### Meten

* Proces waarbij getallen worden toegekend aan variabelen

### Census

* Meten van variabele voor ieder eenheid v/d populatie

### Steekproef

* Deelverzameling v/d eenheden van een populatie
* Moet representatief zijn voor de volledige populatie

### Statistische conclusie

* Schatting, voorspelling of andere generalisatie voor een populatie, gebaseerd op info uit een steekproef

### Betrouwbaarheidsmaat

* Kwantitatieve uitspraak over de mate van onzekerheid van de statistische conclusie

### Vier elementen van de beschrijvende statistiek

1. De populatie (of steekproef)
2. Eén of meerdere variabele(n)
3. Tabellen, grafieken of numerieke hulpmiddelen om een samenvatting te geven
4. Vermeldingen van patronen die in de samenvattingen naar voren komen

### Vijf elementen van de verklarende statistiek

1. De populatie
2. De steekproef
3. Eén of meerdere variabele(n)
4. De conclusie over de populatie, gebaseerd op informatie in de steekproef
5. Een betrouwbaarheidsmaat voor de conclusie

## Soorten Gegevens

### Kwantitatieve gegevens

* Meetwaarden die worden geregistreerd op een van nature voorkomende numerieke schaal.
* = numerieke gegevens

#### Discrete gegevens

* Kwantitatieve gegevens waarbij de variabele een eindig aantal verschillende warden kan aannemen (gehele getallen)
* Komen tot stand door telling
* Bv.: Leeftijd, Score, Aantal passagiers

#### Continue gegevens

* Kwantitatieve gegevens waarbij de variabele een oneindig aantal verschillende waarden kan aannemen (reële getallen)
* Tussen 2 opeenvolgende gegevens kunnen oneindig veel andere waarden gelegen zijn.
* Komen tot stand door meting
* Bv.: Lengte, Grootte, Tijd, Temperatuur

### Kwalitatieve gegevens

* Metingen die niet op een natuurlijk voorkomende numerieke schaal kunnen worden gemeten
* Indelen in categorieën
* = categorische gegevens
* Bv.: Maanden, Geslacht, Lijst v̄ kleuren

#### Nominale gegevens

* Kwalitatieve gegevens waarbinnen er verschillende categorieën aan te duiden vallen
* Geen rangordelijke verschillen tussen categorieën
* Bv.: Man/Vrouw, Provincies, Automerken

#### Ordinale gegevens

* Kwalitatieve gegevens die categorieën bevatten die te rangschikken zijn
* Wel ordening aanwezig
* Bv.: Goed/Matig/Slecht, Goud/Zilver/Brons, Koud/Lauw/Warm/Heet

Diagram

Description automatically generated

## Voorstellen en ordenen van gegevens

### Kwalitatieve gegevens

#### Frequentietabel en staafdiagram

#### Klasse

* Eén v/d categorieën waarin KG (kwalitatieve gegevens) kunnen worden ingedeeld

#### Absolute frequentie

* Aantal waarnemingen in verzameling die tot bepaalde klasse behoren

#### Relatieve frequentie

* Percentage

### Continue gegevens

#### Groeperen van gegevens

* Groeperen in Stengel-blad diagram
* Groeperen in klassen
  + Frequentietabel
    - Klassen, AF (ni), CAF (Ni), RF (fi), CRF (Fi)
  + Histogram
  + Cumulatief frequentiepolygoon

## Centrummaten

### Sommatienotatie

* A picture containing clock, watch

  Description automatically generatedn = aantal variabelen
* xi = eerste variabele
* Bv.:

Text

Description automatically generated with low confidence

### Notaties en definities van centrummaten

#### Centrum

* De neiging v/d gegevens om zich rond een bepaalde waarde te concentreren
  + Table

    Description automatically generatedGemiddelde
  + Modus
  + Mediaan

#### A picture containing text Description automatically generatedGemiddelde

* Beïnvloed door extremen

##### Gemiddelde als evenwichtspunt

* A picture containing watch, clock

  Description automatically generatede = squared error
* m = waarde
* e is minimaal wanneer m = x̄

##### Gemiddelde van discrete gegevens

* niet altijd een variabele
* Bv.: gemiddelde aantal kinderen = 1,58

##### Gemiddelde van continue gegevens

* Data gegroepeerd in klassen
* Exacte waarde = verloren
* Klassenmidden = mi
* mi =

#### Modus

* Meest voorkomende waarde
* Maat voor centrale tendens
* Niet beïnvloedt door extremen
* Geen of meerdere modi mogelijk
* Zowel kwal. Als kwan. Data

#### Mediaan

* Middelste waarde v/e geordende rij
  + Maat voor centrale tendens
  + n = oneven: middelste waarde
  + n = even: gemiddelde van 2 middelste waarden
  + positie v/d mediaan in een rij =
  + niet beïnvloedt door extremen

##### Mediaan van gegroepeerde gegevens

* xme = L + (pN – NL)\*(C/ni)
* L = ondergrens mediale klasse
* P = 0.5 =
* NL = Cumulatieve freq. tot benendengrens
* Ni = AF v/d mediale klasse
* C = Klassenbreedte
* N = totaal aantal waarnemingen

### Overzicht toepasbaarheid v/d centrummaten

Table

Description automatically generated

## Scheefheid

* Kurtosis / Skew

### Soorten scheefheid

* Diagram

  Description automatically generatedLinksscheef
  + X̄ ≤ xme
* Diagram

  Description automatically generatedSymmetrisch
  + X̄ = xme
* Diagram

  Description automatically generatedRechtsscheef
  + X̄ ≥ xme

# Sessie 2 – Spreidingsmaten

## Spreidingsmaten

### Range

* Range = Variatiebreedte = bereik
* Xmax – Xmin
* Gevoelig voor uitschieters

### Interkwartielafstand (IQR)

* Q1 = 25ste
* Q2 = 50ste = [mediaan](#_Mediaan)
* Q3 = 75ste
* IQR = Q3 – Q1

### Boxplot

* Grafische voorstelling v/d spreiding v/d date op basis van 5 kerngetallen:
  + Diagram

    Description automatically generatedMinimum
  + Maximum
  + Median
  + Q1
  + Q3
* Uitschieters = waarden kleinder dan Q1 – (1.5\*IQR) of groter dan Q3 + (1.5\*IQR)
* Staarten passing zich aan zodat lengte max 1.5\*IQR is
* Linkervak (Q1 – Q2) groter = linksscheef
* Rechtervak (Q2 – Q3) groter = rechtsscheef

### Variantie en standaardafwijking

#### Steekproefvariantie S2

* S2 voor steekproef van *n* meetwaarden is gelijk aan de som van de gekwadrateerde afwijkingen van het gemiddelde, gedeeld door (n-1)

#### A picture containing text, clock Description automatically generated

#### Populatievariantie σ2

* A picture containing text

  Description automatically generatedIdem, gedeeld door N (ipv n-1)

#### Standaardafwijking S van een steekproef

* Positieve wortel uit steekproefvariantie S2

#### Standaardafwijking σ van een populatie

* Positieve wortel uit steekproefvariantie σ2

### Variatiecoëfficiënt V

* Dimentieloos getal
* Verhouding STD / Gemiddelde
* V = S/x̄
* V < 5% = zeer kleine spreiding
* 5% < V < 10% = kleine spreiding
* V > 50% = zeer grote spreiding

### Spreidingsmaten vergelijken

* Range
  + Weinig rekenwerk
  + Slechts 2 gegevens
  + Zeer gevoelig voor uitschieters
* IQR
  + Relatief weinig rekenwerk
  + Enkel spreiding middelste groep gegevens
  + Niet gevoelig voor uitschieters
* Variantie/standaardafwijking (STD)
  + Veel rekenwerk
  + Alle gegevens
  + Enkel gevoelig bij kleine n of N

## Interpretatie van de variantie en STD

### Theorema van Chebyshev

### Empirische regel bij symmetrische heuvelvormige verdelingen (normaalverdelingen)

* 68% v/d waarden liggen tussen x̄-S en x̄+S (µ̄-σ en µ̄+σ bij populaties)
* 95% v/d waarden liggen tussen x̄-2S en x̄+2S (µ̄-2σ en µ̄+2σ bij populaties)
* 99.7% v/d waarden liggen tussen x̄-3S en x̄+3S (µ̄-3σ en µ̄+3σ bij populaties)Chart, histogram

  Description automatically generated

### Extra betekenissen van de variantie

* Maat voor onzekerheid: groter variantie = moeilijker voorspellen
* Drukt hoeveelheid ruis/fouten op een signaal uit

## Relatieve maten

### Percentiel

* Voor verzameling van *n* meetwaarden, is het pe percentiel een getal zodat p% van de meetwaarden onder het pe percentiel valt, en (100-p)% erboven valt.
* Cumulatieve proportie
* Q1 = 25ste percentiel
* Mediaan = 50ste percentiel
* Q3 = 75ste percentiel
* Bv.:

Score = 420/500. Met die score in 67ste percentiel. 320 deelnemers

67% scoort minder dan jou = 0.67\*320 = 214 mensen

(100-67)% = 33% scoort meer = 0.33\*320 = 106 mensen

### Z-score

* Populaire maat om relatieve positie van een waarde tov v/d rest v/d data uit te drukken
* Vertegenwoordigt de afstand tussen een gegeven meetwaarde x en het gemiddeld
* Uitgedrukt in aantal standaardafwijkingen
* Steekproef :
* Populatie:
* Bv.:

200 arbeiders. X̄ inkomen = 24000. STD S=2000. Persoon a inkomen = 22000

Z-score van a: STD

* Eigenschappen
  + Relatieve maat voor de meetwaarde
  + 68%: z-score tussen -1 en +1
  + 95%: z-score tussen -2 en +2
  + 99.7%: z-score tussen -3 en +3
* Uitschieters detecteren
  + Z-score groter dan |3|

## Scheefheid en Kurtosis

### [Scheefheid](#_Scheefheid)

* Maat die aangeeft of verdeling links- of rechtsscheef is
* Galton skewness
* SKB = 0 Symmetrisch
* SKB > 0 Rechtsscheef
* SKB < 0 Linksscheef

### Kurtosis

* Maat voor piekvormigheid van een verdeling
* K = 3 Normale verdeling (mesokurtic)
* K > 3 Scherpe piek (leptokurtic)
* K < 3 Stompe piek (platykurtic)

#### Excess kurtosis Ke

* Maat voor piekvormigheid met als referentie de normaalverdeling
* Ke = K – 3
* Ke = 0 Normale verdeling (mesokurtic)
* Ke > 0 Scherpe piek (leptokurtic)
* Ke < 0 Stompe piek (platykurtic)
* =KURT() in excel

## Lineaire transformaties

* Een bewerking waarbij elke waarde x herrekend wordt naar y volgens y = ax+b
* Bewerkingen met wortels, machten, logaritmes, enz. ≠ lineair

### Effect op rekenkundig gemiddelde

* Ȳ = ax̄+b

### Effect op variantie S2 en STD S

* S2y = a2S2x
* Sy = |a|Sx

# Sessie 3 – Discrete kansverdelingen

## Begrippen

### Stochastische variabelen (SV)

* Een variabele die numerieke waarden aanneemt die horen bij de uitkomsten van een toevalsexperiment, waarbij één waarde aan elke uitkomst wordt toegekend.
* = kansvariabele
* Stochastisch: van het toeval afhankelijk
* Bv.:
  + Tellen v/h aantal klanten die geldautomaat gebruiken op 1 dag

#### Discrete stochastische variabele (DSV)

* Telbaar (eindig of oneindig)
* Bv.:
  + Exp.: 100 verkoopspogingen Tijd tussen aankomst klanten
  + SV: aantal verkopen Aantal minuten, uren, seconden
  + Mogelijke uitkomsten: 0-100 0.4, 2, … ∞

## Kansverdeling van discrete stochastische variabelen

### Discrete kansverdeling

* Grafiek, tabel of formule die kansen vastlegt die bij mogelijke waarden die de SV kan aannemen horen
* Bv.: exp waarbij 2 muntstukken worden gegooid, tel aantal keren kop
  + MM: 0, KM: 1, MK: 1, KK: 2
  + P(x=0)=p(MM)=1/4
  + P(x=1)=p(MK)+p(KM)=1/4+1/4=1/2
  + P(x=0)=p(KK)=1/4

#### Voorwaarden

* p(x) ≥ 0 voor alle waarden van x
* ; totale kans van alle mogelijke uitkomsten is 1 = 100%

### Overzicht kerngetallen

#### Verwachtingswaarden (gemiddeld)

* Vb.:
  + winst (x): 290 euro -9710 euro
  + Kans: 0.999 0.001
  + µ = E(290)(0.999)+(-9710)(0.001) = 280 euro

#### Variantie

#### Standaardafwijking

* Table

  Description automatically generatedσ =

Vb.: kansspel met dobbelsteen

µ = E(x) =

µ = 2\* =

σ2 = 3.81

σ =

## Binomiale verdeling

### Eigenschappen van de binomiale verdeling

* Bestaat uit n identieke deelexperiementen
* Slechts 2 mogelijke uitkomsten: S = Succes, M = Mislukking
* Kans op uitkomst S is voor ieder deelexperiment even groot
* Kans op succes: p
* Kans op mislukking: q = 1-p
* Deelexperimenten zijn onafhankelijk
* Binomiale SV x is het aantal keren S in n deelexp.
* Vb.:
  + Kans dat er in een gezin v 5 kinderen 4 meisjes zijn
  + Kans dat er tussen 40 producten 3 defecte zitten
  + Kans dat 4 op 10 klanten iets koopt

### Excel functies

#### Kans berekenen met X

=BINOM.DIST(x;n;p;False/True)

False: voor exact aantal

True: aantal of minder (als aantal of meer: 1-BINOM.DIST)

#### X berekenen met kans

=BINOM.INV(n;p;percentage)

Percentage: minimaal percentage (als maximaal: 1-percentage)

### Rekenen met binomiale kansverdeling

* p = kans op succes in één deelexp.
* q = 1-p
* X = aantal successen in *n* deelexp.
* 5! = 5\*4\*3\*2\*1 = 120 (0! = 1)

### Verwachting en spreidingsmaten v/d binomiale stochastische variabele

* Verwachting: µ = np
* Variantie: σ2 = npq
* Standaardafwijking σ =

Vb.: Een machine die onderdelen maakt voor auto’s functioneert niet goed en levert 1/10 defecte producten. De volgorde is willekeurig. Bepaal de kans dat van de 5 volgende onderdelen er 3 defect zijn.

X aantal defecte onderdelen in n = 5 deelexp.

p = 0.1 en q = 1-p = 0.9

met x = 3 bekomen we:

De kans is 0.81%

In excel: =BINOM.DIST(3;5;0.1;False)

## Poissonverdeling

### Eigenschappen van de poissonverdeling

* Beschrijven van het aantal gebeurtenissen in een bepaalde tijdsperiode, oppervlak of in bepaald volume (of gewicht, afstand of andere meeteenheid)
* De kans dat een gebeurtenis voorkomt is even groot voor alle tijdsperiodes, oppervlakken,… van gelijke grootte
* Gebeurtenissen zijn onafhankelijk van het aantal dat in andere disjunct eenheden voorkomt
* Aantal gebeurtenissen in elke eenheid: λ (=µ = σ2)

### Excel functies

#### Kans berekenen met X

=POISSON.DIST(x; λ; True/False)

False: voor exact aantal

True: aantal of minder (als aantal of meer: 1-POISSON.DIST)

### Rekenen met de poissonverdeling

(met x = 0, 1, 2, …)

p(x) = de kans dat x gebeurtenissen voorkomen in het tijdsinterval

e = 2.71828 (constante van Euler)

Vb.: per maand gebeuren in de luchtvaart gemiddeld 4.4 ongelukken. Veronderstel dat de kansverdeling v/h aantal ongelukken x in een bepaalde maand kan worden benaderd door een Poissonverdeling.

1. Bepaal verwachtingswaarde en STD van x

µ = λ = 4.4 σ2 = λ = 4.4

1. Bereken de kans dat er in 1 maand geen enkel ongeluk gebeurd

=> (of 1.23%)

1. Bereken de kans dat er precies 1 ongeval gebeurd

=> (of 5.4%)

1. Bereken de kans dat er meer dan 6 ongevallen gebeuren

(of 15.6%)

# Sessie 4 – Continue kansverdelingen

## Eigenschappen

1. Is een vloeiende kromme aangeduid met f(x)
2. Kromme wordt kansdichtheid, dichtheidsfunctie of dichtheid genoemd
3. Totale opp. onder kromme = 1
4. Kans dat een continue variabele x een waarde aanneemt tussen a en b P(a<x<b) is gelijk aan de opp. onder de kromme tussen de grenzen a en b en kan met integraalrekening worden gezocht. (enkel via software of tabellen te kunnen/kennen)
5. P(x = a) = 0 (opp. onder een enkel punt is 0)

## Normale verdeling

### Eigenschappen v/d normale verdeling

1. Gaussiaanse verdeling (Gaussian / Normal distribution)
2. 1 v/d meest voorkomende kansverdelingen
3. Dichtheid = klokvormig
4. Symmetrisch rond µ
5. Spreiding bepaald door STD σ

### Invloed v/h gemiddelde µ



### Invloed v/d STD σ



### Voorbeelden van normale verdelingen

* Lengtes van personen
* Zwangerschapsduur
* Levensduur van een lamp
* Schoenmaten
* Bloeddruk
* IQ
* …

### Rekenen met normale verdeling

#### Kansverdeling voor een normaal verdeelde variabele x

µ = verwachting v/d normaal verdeelde variabele x

σ = STD

π = 3.1415…

e = 2.71828…

#### Empirische regel bij symmetrische heuvelvormige verdelingen

* 68% v/d waarden liggen tussen x̄-S en x̄+S (µ̄-σ en µ̄+σ bij populaties)
* 95% v/d waarden liggen tussen x̄-2S en x̄+2S (µ̄-2σ en µ̄+2σ bij populaties)
* 99.7% v/d waarden liggen tussen x̄-3S en x̄+3S (µ̄-3σ en µ̄+3σ bij populaties)Chart, histogram

  Description automatically generated

#### Herkennen v/d normale verdeling

NIET ALLES IS NORMAAL VERDEELD

* Kijk naar histogram. Klokvormig en symmetrisch?
* Bepaal centrumwaarden. Zijn mediaan, modus en gemiddelde (ongeveer) gelijk?
* Empirische regel

#### Standaard normale verdeling

* µ = 0 & σ = 1
* Standaardnormaal verdeelde variabele : z

Vb.:

P(z<-0.67)

P(z<-0.67) = 0.2514 = 25.143%

### Rekenen met niet-standaard normale verdelingen

#### Standaardiseren van niet-standaardnormale verdelingen

* Bij een normale verdeling waarbij µ≠0 of σ≠1 => standaardiseren via z-score
* Z-waarde geeft aantal std’s tussen x en µ
* Na standaardisatie gelde rekenregels van standaard normale verdelingen

#### Voorbeelden

P(x<7) met µ = 9 en σ = 4

(via tabel)

P(z<-0.50) = 0.30854 = 30.854%

**x-waarde vinden op basis van kans**

IQ: normaal verdeeld; µ=100; σ=15

Bepaal minimale IQ van de intelligentste 10%

(via tabel)

z-score 10%

1– 0.1= 0.9

=> z ≈ 1.29 => x = z\*σ+µ = 1.29\*15+100 = 119.35

# Excel en Python functies overzicht

## Excel

|  |  |
| --- | --- |
| =Average | Gemiddelde berekenen |
| =Count | Aantal waarden tellen |
| =Min | Kleinste waarde |
| =Max | Grootste waarde |
| =Stdev.S | Standaardafwijking berekenen |
| =Median | Mediaan |
| =Mode | Modus |
| =Var.S | Variantie |
| =Quartile | Kwartiel (1, 2 of 3) |
| =Skew | Scheefheid |
| =Kurt | Kurtosis |
| =Frequency | Absolute frequentie berekenen |
| =IF | Filteren van bepaalde data  Bv.: =MIN(IF(A2:A25=”female”;B2:B25)   * De kleinste waarde uit alle waarden uit de kolom B2:B25 die in de kolom A2:A25 de waarde “female” hebben |

|  |  |
| --- | --- |
| Klassenbreedte |  |
| Range |  |
| Relatieve frequentie | =AF/aantal |

## Python

|  |  |
| --- | --- |
| df = pd.read\_csv(‘heart.csv’,header = 0) | De dataset met naam ‘heart.csv’ inlezen en aangeven dat deze al een hoofding heeft op positie 0 |
| df.head() | Eerste 5 rijen printen (checken of correct ingelezen) |
| df.describe() | Volledige dataset printen |
| df.shape() | Toont aantal rijen en aantal kolommen |
| df.shape[0] | Toont aantal rijen |
| df.shape[1] | Toont aantal kolommen |
| df[“kolomnaam”] = df.kolomnaam | Geeft enkel die kolom |
| df.kolomnaam.value\_counts() | Geeft de frequentie van alle verschillende waarden |
| df.kolomnaam.max() | Maximum waarde in die kolom |
| df.kolomnaam.min() | Minimum waarde in die kolom |
| df.kolomnaam.mean() | Gemiddelde waarde in die kolom |
| df.kolomnaam.median() | Mediaan in die kolom |
| df.kolomnaam.var() | Variantie |
| df.kolomnaam.quantile(0.25 of 0.75) | Q1 of Q2 |
| df.kolomnaam.skew() | Scheefheid |
| df.kolomnaam.kurt() | Kurtosis |
| df.kolomnaam.std() | Standaardafwijking |
| stats.binom.pmf(x,n,p) | Binomiaal exact aantal |
| stats.binom.cdf(x,n,p) | Binomiaal minder dan |
| stats.binom.sf(x, λ) | Binomiaal meer dan |
| stats.poisson.pmf(x,λ) | Poisson exact aantal |
| stats.poisson.cdf(x, λ) | Poisson minder dan |
| stats.poisson.sf(x, λ) | Poisson meer an |

### Histogram maken

1. klassenbreedte = range/np.sqrt(df.kolomnaam.count())
2. minimum = df.kolomnaam.min()
3. maximum = df.age.max()
4. bins : np.arange(start=minimum, stop=maximum+round(klassenbreedte), step=round(klassenbreedte))
5. sns.distplot(x=df.kolomnaam, bins=bins, kde:False)
6. plt.title(“Histogram)

### Boxplot maken

1. sns.boxplot(data=df, x=’kolomnaam’, whis=1.5)
2. plt.title(‘Boxplot’)

2 boxplots in 1 grafiek:

1. sns.boxplot(data=df, x=’kolom\_1’, y=’kolom\_2’, whis=1.5)
2. plt.title(“boxplot van kolom 1 en kolom 2”)
3. plt.xticks([0, 1], [“onderverdeling 1”, ‘”onderverdeling 2”]

### Andere grafieken

**Countplot**

Sns/catplot(x=’kolomnaam’, kind=’count’, data=df)