Data Analyse en Statistiek

Hella Snoek & Marthe Schut

Inhoud

#Module 1

Hoofdstuk 1

Kanstheorie

In dit hoofdstuk leren we over kanstheorie en kansdichtheidsfuncties. Kanstheorie speelt een belangrijke rol in het begrijpen en bepalen van meetonzekerheden. Zoals in het hoofdstuk over meetonzekerheden is uitgelegd kunnen meetonzekerheden verschillende oorzaken hebben. Bij elk van die oorzaken hoort een bepaalde waarschijnlijkheidsverdeling en deze zijn verbonden aan kans processen.

Vaak willen we metingen gebruiken om voorspellingen te doen of hypotheses te toetsen. Als we een serie meetgegevens hiervoor willen gebruiken, dan is het belangrijk om te weten wat de meetonzekerheden zijn. Deze kunnen we vervolgens gebruiken om te kijken hoe goed ze passen bij een weerpatroon of hoe goed ze een theorie bevestigen of juist weerleggen.

Om die stap later te kunnen maken, moeten we eerst meer leren over kanstheorie en hierna over kansdichtheidsfuncties. In dit hoofdstuk maken we daar een begin mee.

1.1 Definitie van Kans

Waarschijnlijk is iedereen wel bekend met het concept van kans. We gebruiken het vaak. Wat is de kans dat het regent? Wat is de kans om de loterij te winnen?

Wiskundig is een kans gedefinieerd als een getal tussen de 0 en de 1 dat aangeeft hoe waarschijnlijk het is dat een bepaalde gebeurtenis zal plaatsvinden. Een kans van 1 zegt dat het **zeker** zal gebeuren en een kans van 0 dat het **zeker niet** zal gebeuren. Een kans van 0.5 geeft aan dat in 50% van de gevallen de gebeurtenis zal plaatsvinden.

Voorbeeld:

Voorbeeld We kijken naar een dobbelsteen. Wat is de kans dat je een 4 gooit als je de dobbelsteen 1 keer gooit? Voor een normale dobbelsteen kunnen we deze kans uitrekenen met behulp van de volgende formule:

$P(\text{uitkomst is 4}) = \frac{\text{aantal uitkomsten met een 4}}{\text{totaal aantal uitkomsten}} = \frac{1}{6}$

Dit is de kans voor een normale eerlijke dobbelsteen. Met eerlijk bedoelen we hier dat de dobbelsteen niet gemanipuleerd is en dat elk vlak van de dobbelsteen evenveel kans heeft om boven te eindigen.

Stel nu dat we een speciale, waar wel eerlijke, dobbelsteen zouden hebben met de volgende vlakken: {1,2,2,3,4,4}. De mogelijke uitkomsten bij een dobbelsteenworp zijn nu: {1,2,3,4}. Dit noemen we ook de **uitkomstenverzameling** waarbij alle elementen uniek zijn, en dus maar 1 keer voorkomt.

De kans om nu een 4 te gooien is groter dan met een normale eerlijke dobbelsteen, namelijk.

Voorbeeld:

Voorbeeld Als we de kans nu berekenen voor de speciale dobbelsteen met vlakken {1,2,2,3,4,4} dan is de kans om vier te gooien:
$P(\text{uitkomst is 4}) = \frac{\text{aantal uitkomsten met een 4}}{\text{totale aantal uitkomsten}} = \frac{2}{6}$

En stel nu dat we een normale dobbelsteen hebben die gemanipuleerd is? Dan zal de kans om een 4 te gooien anders zijn. Een goede manier om dan de kans te bepalen is met behulp van de **Frequentist** formule:

$$P(4) = lim_{n\to\infty} \frac{\text{uitkomst is 4}}{\text{totaal aantal worpen}} \tag{1.1}$$

De algemene formule voor de **Frequentist definitie** van kans is:

$$P(\text{uitkomst A}) = \lim_{n\to\infty} \frac{\text{uitkomst A}}{n}. \tag{1.2}$$

Waarbij we n metingen hebben verricht.

De Frequentist definitie voor kans is een goede manier om kansen te berekenen. Het kent echter twee grote beperkingen. De eerste is dat we eigenlijk nooit een oneindig aantal metingen kunnen doen. Dit is goed te benaderen door gewoon een heel groot aantal metingen te doen. De tweede beperking is dat niet alle experimenten herhaalbaar zijn.

1.1.1 Frequentist versus Bayesiaanse methode

Het zal je dan misschien niet verbazen dat er nog een andere methode bestaat die wel werkt voor experimenten die niet herhaalbaar zijn of een beperkte statistiek hebben. Deze manier noemen we ook wel de Bayesiaanse (spreek uit: Beej-sie-jaanse) methode (Engels: Bayesian).

De frequentist methode wordt in het algemeen als objectieve methode gezien en de Bayesiaanse methode een subjectieve manier. Het geeft aan wat je denkt dat de waarschijnlijkheid is. Dat klinkt misschien niet erg wetenschappelijk maar in de praktijk is dit misschien wel de meest gebruikte methode. Vooral omdat je hem ook kan gebruiken als het experiment niet herhaalbaar is. De bayesiaanse methode zegt eigenlijk dat je het nooit helemaal zeker kunt stellen wat een kans is. Dat voelt misschien wat gek, maar het enige wat het zegt is dat ook bij een berekende kans waarde er een mate van onzekerheid is. Ook daar is er sprake van een 'meetonzekerheid'.

Voorbeeld:

Een voorbeeld In een wielerronde staat een bergklassieker op het programma van vandaag. De wedstrijd is nog niet gestart. Er staan twee sterke renners, Verstappen en Onana, op de gedeelde eerste plaats van het klassement en de voorsprong met de derde wielrenner is meer dan 20 minuten. Het lijkt dus waarschijnlijk dat aan het einde van de dag Verstappen of Onana op de eerste plaats in het klassement zal staan. Op bergetappes wint Onana 9 van de 10 keer met een flinke voorsprong van Verstappen. Wie denk je dat er vandaag wint?

We kunnen het experiment natuurlijk niet herhalen maar het lijkt zeer waarschijnlijk dat Onana aan het einde van de dag op nummer 1 zal eindigen. Hier maken we gebruik van de subjectieve methode van Bayes. Om het te kwantificeren kunnen we misschien zelfs wel zeggen dat de kans 0.9 is.

Maar nu zitten we aan het ontbijt en we zien dat Onana geen hap door zijn keel krijgt. Hij is duidelijk erg ziek. Verstappen daarentegen ziet er fris en sterk uit.
Hoe waarschijnlijk denk je nu dat het is dat Onana zal winnen?

Het lijkt nu toch een stuk minder waarschijnlijk dat Onana zal winnen. Misschien schat je nu de kansen lager in dan de 0.9 waarmee je begon. Misschien heb je zelfs wel informatie uit het verleden waaruit je weet hoeveel langzamer renners zijn als ze er zo ziek uitzien als Onana. Wat voor impact dat heeft op hun performance. Dan zouden we ons kans van 0.9 kunnen 'updaten' met de nieuwe informatie. Dat is typisch een Bayesiaanse methode om kansen uit te rekenen.

Beide methodes worden dus gebruikt, maar de Bayesiaanse methode, of zelfs een hybride methode vindt vooral zijn toepassing in heel complexe modellen en voorspellingen. In dit vak zullen we echter vooral werken met de frequentist methode. Wat in elk geval belangrijk is, is om altijd heel precies te vermelden wat de voorwaardes zijn geweest waaronder de kans is uitgerekend. Ook bij de frequentist methode!

1.2 Rekenen met kansen

Er zijn een paar basisregels waar kansen aan voldoen.

1. **Behoud van kans:**Een gebeurtenis, A, kan plaatsvinden, of het kan niet plaatsvinden. De kans is behouden en dat betekent dat:

$$P(A) + P(\text{niet A}) = 1 \tag{1.3}$$

2. **Complementregel:** Een direct gevolg hiervan is dat $P(\text{niet A})$ het complement is van $P(A)$ ofwel:

$$P(\text{niet A}) = 1 - P(A). \tag{1.4}$$

3. Als de uitkomst B *bestaat* dan geldt:

$$0 < P(B) \leq 1. \tag{1.5}$$

   Een kans moet dus altijd groter zijn dan nul voor alle elementen in de uitkomstenverzameling.

4. **De *of* Regel**: Als de uitkomsten A en B *wederzijds uitsluitend* zijn, ofwel als A plaats vindt, dan kan B nooit plaats vinden, dan geldt:

$$P(A \text{ of } B) \equiv P(A \cup B) = P(A) + P(B). \tag{1.6}$$

   We mogen in dit geval de kansen dus optellen.

5. **De *en* regel**: Als de uitkomsten A en B onafhankelijk zijn, dus als je A een uitkomst is dan zegt dat niets over de kans op B, dan geldt:

$$P(A \text{ en } B) = P(A) \cdot P(B). \tag{1.7}$$

We gaan voor elk van deze regels een voorbeeld geven. We kijken hiervoor naar een kaartendek. De uitkomstenverzameling van een kaartendek is:

1♡,2♡,3♡,4♡,5♡,6♡,7♡,8♡,9♡,H♡,D♡,K♡,A♡,
1♢,2♢,3♢,4♢,5♢,6♢,7♢,8♢,9♢,H♢,D♢,K♢,A♢,*
1♠,2♠,3♠,4♠,5♠,6♠,7♠,8♠,9♠,H♠,D♠,K♠,A♠,
1♣,2♣,3♣,4♣,5♣,6♣,7♣,8♣,9♣,H♣,D♣,K♣,A♣

Dit zijn in totaal 52 kaarten verdeeld over 2 kleuren: rood en zwart. We trekken in de volgende voorbeelden steeds 1 kaart.

Voorbeeld:

Voorbeeld 1 - behoud van kans/complement regel:
* De kans om een harten 5 uit een dek kaarten te trekken is precies: P(5♡)= *1/52.* De kans om een *andere kaart dan een harten 5* te trekken is gelijk aan: 1-P(5♡) = *1-1/52 = 51/52.*
De kans om een rode kaart te trekken is precies 26/52 = 1/2 en is precies gelijk aan de kans om een zwarte kaart te trekken (1-1/2 = 1/2).

Voorbeeld:

Voorbeeld 2 - groter dan nul:
* Voor elke kaart in het dek is er een kans dat je hem trekt.

Voorbeeld:

Voorbeeld 3 - de of-regel:
* De kans dat je een 3 of een 5 trekt is gelijk aan P(3)+(P(5) = 1/13+1/13 = 2/13.
* De kans dat je een 3 of een rode kaart trekt kunnen we niet zomaar optellen. Er bestaan ook rode kaarten met een 3.

Voorbeeld:

Voorbeeld 4 - de en-regel:
* De kans dat je een 3 trekt die ook een rode kaart is kunnen we uitrekenen met:

$$P(\text{rood en } 3) = P(\text{rood}) \cdot P(3) = 1/2 \cdot 4/52 = 2/52 \tag{1.8}$$

Er zijn maar twee rode 3 kaarten in het dek, dus dat klopt. Er zijn evenveel rode drie kaarten als zwarte drie kaarten en daarom mag je ze in dit geval vermenigvuldigen. De uitkomsten zijn onafhankelijk.
onafhankelijk. Als je een 9♡ trekt, zegt dat al direct iets over de kans dat deze kaart ook een A♣ is (die is namelijk gereduceerd tot 0).

Hoofdstuk 2

Kansdichtheidsfuncties

We gaan nu kijken naar kansverdelingen. In het voorbeeld van de simpele dobbelsteen zou je kunnen kijken hoe de kansen verdeeld zijn over de verschillende uitkomsten. Voor een normale dobbelsteen is dit misschien een beetje saai, voor elke uitkomst verwacht je een andere waarde. Voor de speciale dobbelsteen die we eerder beschreven ziet het er al wat interessanter uit.

Om wat over kansverdelingen te kunnen schrijven moeten we eerst weten wat stochasten zijn. Daarna introduceren we enkele veelgebruikte kansdichtheidsverdelingen.

2.1 Wat is een stochast?

Een **stochast** is een variabele waarvan de waarde van een kans proces afhangt. Bijvoorbeeld de uitkomst van het trekken van een kaart, dan is het getrokken kaart (de uitkomst van de trekking) een stochast. Je weet van tevoren niet welke kaart je gaat trekken en daarom is de uitkomst *stochastisch*. Of als je een met een dobbelsteen gooit dan is de uitkomst van de worp een stochast. Het Engelse woord (random variable) is misschien bekender.

2.2 Kansdichtheidsfuncties

Stochasten zijn een handig middel bij het beschrijven van experimenten. We gaan hieronder een aantal vaak voorkomende distributies van stochastische variabelen bekijken. De distributies laten zien wat de kans is dat een bepaalde stochastische waarde wordt gevonden. Het is dus een verdeling van kansen. Deze verdelingen noemen we **kansdichtheidsfuncties** (Engels: probability density function of PDF). Een kansdichtheidsfunctie, $f(x)$, zegt dat de kans dat een variabele x gevonden wordt in een gebied $[x, x + dx]$ gelijk is aan $f(x)dx$.
De kans dat we x terugvinden in een interval $[a, b]$ is gelijk aan:

$$P(a \leq x \leq b) = \int_a^b f(x)dx. \tag{2.1}$$

Er zijn **twee belangrijke voorwaardes** aan een kansdichtheidsfuncties die je misschien bekend zullen voorkomen:

uitkomstengebied.
2. De kansdichtheidsdistributie moet genormaliseerd zijn op 1.
In formule notatie: $f(x) \geq 0$ en $\int_{-\infty}^{\infty} f(x)dx = 1$.

Wellicht komt dit allemaal wat abstract over en helpt het om wat concrete voorbeelden te zien. Hieronder definiëren we vier belangrijke kansdichtheidsfuncties (ook wel PDFs). Er zijn veel meer kansdichtheidsfuncties gedefinieerd, kijk bijvoorbeeld maar eens naar deze lijst op Wikipedia.

Voor we gaan kijken naar de voorbeelden is het handig om uit te leggen hoe we de verwachtingswaarde en de standaarddeviatie kunnen uitrekenen voor kansdichtheidsfuncties. De definities hiervan heb je gezien in het hoofdstuk Basisbegrippen, voor dichtheidsfuncties zien de formules er net iets anders uit dan voor datasets.

2.3 Verwachtingswaarde en standaarddeviatie

Voor **discrete** verdelingen gelden de volgende vergelijkingen:

- de verwachtingswaarde: $\mu = E(x) = \sum_{i=1}^{N} x_i P(x_i)$,
- de variantie: $\sigma^2 = \sum_{i=1}^{N} (x_i - E(x))^2 P(x_i)$.

Voor **continue** verdelingen maak je gebruik van de volgende vergelijkingen:

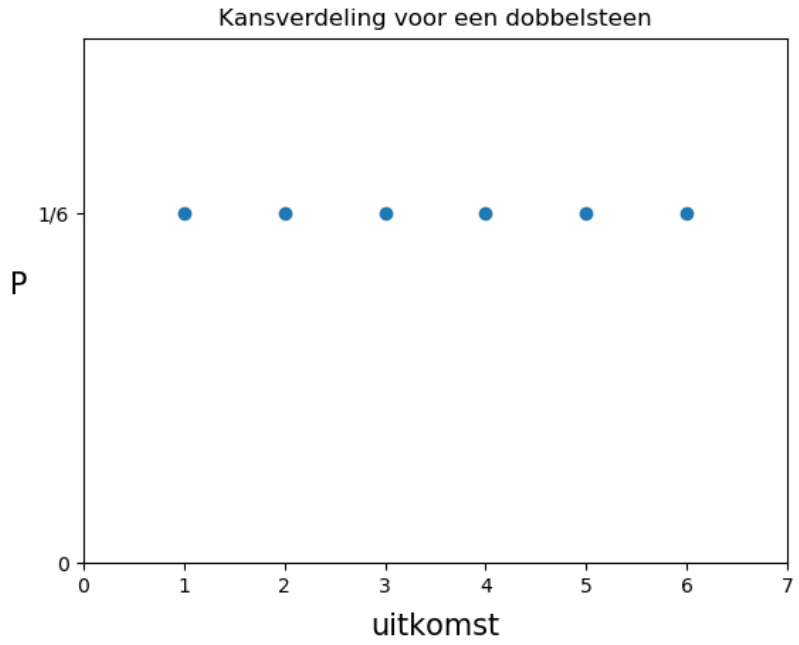
- de verwachtingswaarde: $\mu = E(x) = \int_{-\infty}^{\infty} x f(x)dx$,
- de variantie: $\sigma^2 = E(x^2) - E(x)^2 = \int_{-\infty}^{\infty} (x - E(x))^2 f(x)dx$.

NB Herinner je nog het verschil tussen parameters (voor de kenmerken van een populatie) en statistieken (voor de kenmerken van een steekproef). Afhankelijk van wat we beschrijven zijn verschillende schrijfwijze voor het gemiddelde $\mu, < x >$ en $E(x)$. Het symbool μ is meestal voorbehouden aan het gemiddelde van de populatie, dat wil zeggen het *echte* gemiddelde. Het gemiddelde van de steekproef is $< x >$, je hoopt dus dat die dicht bij het populatie gemiddelde μ ligt. De verwachtingswaarde $E(x)$ is de waarde die je verwacht te gaan meten. Deze kan je met simulaties benaderen. De verschillen worden pas echt duidelijk als je er al een tijdje mee werkt. We zullen het niet fout rekenen als je een vergissing maakt in de notatie, maar we proberen het hier wel netjes op te schrijven. In deze vergelijkingen is het in elk geval ook gewoon handiger om $E(x)$ of $< x >$ te schrijven. $E(x)^2$ is, net als $< x >^2$, het kwadraat van de verwachtingswaarde van x. $E(x^2)$ is, net als $< x^2 >$ de verwachtingswaarde van x^2. De kansdichtheidsverdeling.

2.4 Bekende kansdichtheidsfuncties

2.4.1 Uniform

De uniforme distributie is een vlakke kansverdeling. De kans op elk deel van de uitkomstenverzameling is gelijk. We hebben hier al een paar voorbeelden van gezien. Bijvoorbeeld bij de eerlijke dobbelsteen waarbij de kans op elk van de 6 uitkomsten precies gelijk is. De uitkomsten van een dobbelsteen zijn discreet. Voor **discrete uniforme** verdelingen van stochastische waarden kunnen we schrijven dat de kans op uitkomst van stochast i, $P(i)$, gevonden kan worden met de relatie: $P(i) = 1/N$.
Waarbij N de hoeveelheid mogelijke uitkomsten is. Dit ziet er grafisch als volgt uit:



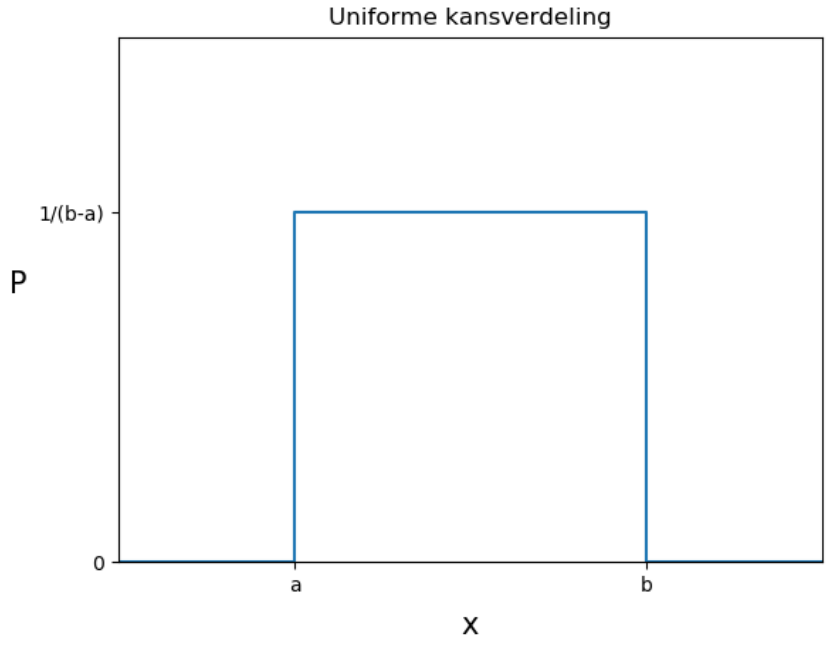
Een algemene formule voor een **continue uniforme** verdeling is:

$$f(x; a, b) = \frac{1}{b-a} \quad \text{voor} \quad a \leq x \leq b.$$

Hierbij is $f(x)$ de kans dat je de waarde x vindt. De stochast is hier dus x. Hieronder zie je hoe de uniforme verdeling eruit ziet voor een continue verdeling:

De verwachtingswaarde en de standaarddeviatie van de uniforme verdeling zijn $E(x) = (a+b)/2$ en $\sigma = (b-a)/\sqrt{12}$.

De **verwachtingswaarde** kunnen we uitrekenen met behulp van de algemene formule:



$$\begin{aligned} E(x) &= \int_{-\infty}^{\infty} xf(x)dx = \int_a^b x \cdot \frac{1}{b-a}dx \\ &= \frac{1}{2}\frac{1}{(b-a)}x^2\Big|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}. \end{aligned} \tag{2.2}$$

De **standaarddeviatie** berekenen we met de formule:

$$\begin{aligned} \sigma^2 &= \int_{-\infty}^{\infty} (x-E(x))^2 f(x)dx = \int_a^b \left(x - \frac{a+b}{2}\right)^2 \cdot \frac{1}{b-a}dx \\ &= \frac{1}{12} \cdot \frac{(b-a)^3}{b-a} = \frac{(b-a)^2}{12}. \end{aligned} \tag{2.3}$$

Dit geeft de vergelijking voor de standaarddeviatie: $\sigma = \frac{(b-a)}{\sqrt{12}}$.

2.4.2 Binomiaal

Om de binomiale verdelingsfunctie uit te leggen beginnen we eerst met het Bernoulli-experiment. Dit is een experiment met maar twee uitkomsten, 'succes' en 'mislukking'. De kans op succes is p en de kans op mislukking q, dan is dus $q = 1 - p$.

Als we precies n onafhankelijke Bernoulli experimenten uitvoeren dan is de kans op een totaal aantal malen succes uit deze n experiment gedefinieerd als k. Dit wordt beschreven door de binomiale verdeling:

$$P(k;n,p) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} \equiv \frac{n!}{k!(n-k)!} p^k q^{n-k}.$$

Het gemiddelde en de standaarddeviatie van de Binomiale verdeling zijn:

$\sigma = \sqrt{npq}$.

Voorbeeld:

Voorbeeld Stel dat we een oneindige grote verzameling knikkers hebben waarvan 30% gele knikkers, alle andere knikkers zijn rood gekleurd. Als we een enkele knikker trekken hebben we dus precies 30% kans ($p = 0.3$) dat dit een gele knikker is.
Als we twee knikkers trekken hebben we een kans van $0.3 \cdot 0.3 = 0.09$ dat we precies twee gele knikkers hebben getrokken. Immers, omdat de verzameling oneindig groot is, heeft de eerste trekking geen invloed op de tweede trekking en zijn de twee trekkingen onafhankelijk. We mogen dus de 'en'-regelgebruiken.
We hebben een kans van $(1 - 0.3 * 0.3) = 0.91$ dat we minstens 1 rode knikker hebben, hier gebruiken we de complement regel.
De kans dat we twee rode knikkers hebben (en dus geen gele knikkers) is $(1 - 0.3) \cdot (1 - 0.3) = 0.49$. We kunnen nu ook redeneren dat de kans dat we 1 gele knikker en 1 rode knikker hebben getrokken precies gelijk is aan $0.91 - 0.49 = 0.42$.

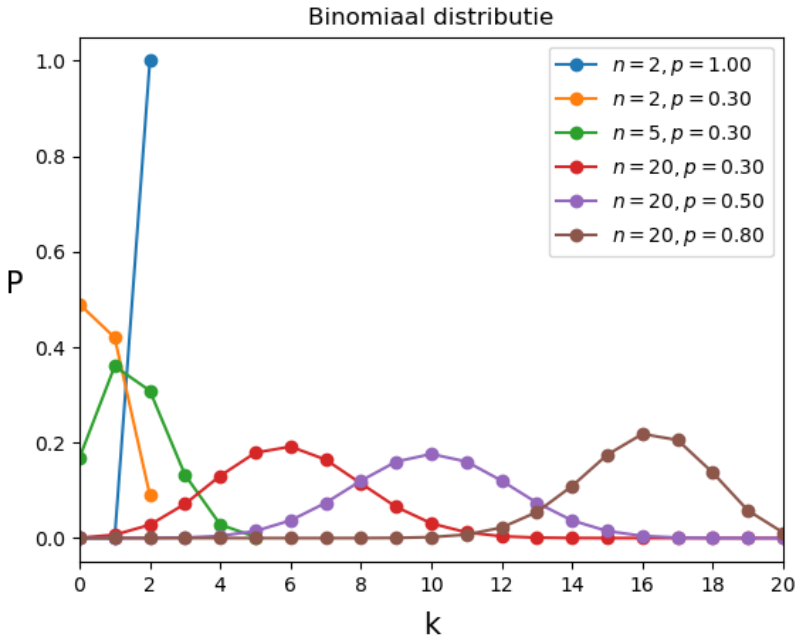
We kunnen deze kansen ook met de Binomiaal vergelijking uitrekenen:
2 trekkingen, 0 gele knikkers: $P(k;n,p) = p(0;2,0.3) = \frac{2!}{(0!\cdot 2!)} 0.3^0 \cdot 0.7^2 = 0.49$
2 trekkingen, 1 gele knikkers: $P(k;n,p) = p(1;2,0.3) = \frac{2!}{1!\cdot 1!} 0.3^1 \cdot 0.7^1 = 0.42$
2 trekkingen, 2 gele knikkers: $P(k;n,p) = p(2;2,0.3) = \frac{2!}{2!\cdot 0!} 0.3^2 \cdot 0.7^0 = 0.09$

Deze kansen staan ook uitgedrukt in de gele lijn in het plaatje hieronder.

De binomiale verdeling is een discrete verdeling. Deze formule kunnen we niet toepassen op fractionele waardes. Dat is ook logisch want het Bernoulli experiment kunnen we niet een fractioneel aantal keer uitvoeren. De kansverdeling is asymmetrisch voor lage waardes van n en wordt voor grotere waardes van n steeds meer symmetrisch.

Hieronder zie je een aantal verdelingen voor de Binomiaal.

Het voorbeeld van daarnet is uitgedrukt in de gele lijn. Kijk ook eens goed naar de blauwe lijn. De kans $p = 1$ zegt dat een de uitkomst altijd succes is. Als je het experiment twee keer uitvoert, zijn ze dus gegarandeerd allebei succesvol. En de kans is 0 dat je maar 1 uit 2 ($n = 2, k = 1$) positieve uitslagen hebt. Dat kan immers ook niet, je kan alleen maar succes hebben, er bestaan geen andere uitslagen van het experiment.



2.4.3 Poisson

De Poisson is discrete verdelingsfunctie die, in veel gevallen, de onzekerheid weergeeft op telexperimenten. Het aantal geobserveerde gebeurtenissen (k) is gerelateerd aan het verwachte aantal gebeurtenissen (λ) via de Poissonverdeling:

$$P(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{2.4}$$

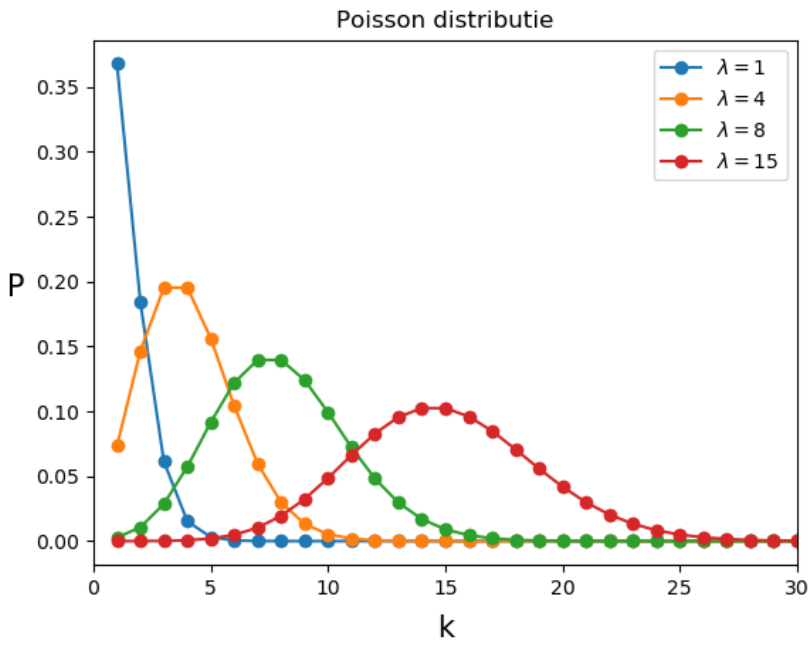
De Poisson kent, in tegenstelling tot de binomiaal dus maar 1 parameter. De verwachtingswaarde van de Poisson vergelijking (het gemiddelde) is λ en de variantie is ook λ. De onzekerheid op een stochast, als deze de Poisson statistiek volgt, is gelijk aan de standaarddeviatie: $\sigma = \sqrt{\text{var}} = \sqrt{\lambda}$.

Het is dus een bijzondere vergelijking! Hieronder zie hoe de Poisson distributie eruit ziet voor verschillende waardes van λ.

De Poisson verdeling is, net als de Binomiaal vergelijking asymmetrisch voor lage waardes van λ en wordt voor steeds meer symmetrisch voor hogere waardes van λ. Dat is ook geen toeval, de Poisson vergelijking is een speciale vorm van de Binomiaal. Als je hier meer over wilt weten kun je dit filmpje bekijken.

2.4.4 Normaal (ofwel Gauss)

Stochastische variabelen zijn normaal verdeeld (ook wel Gaussisch) als ze door de volgende functie worden beschreven:



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}.$$

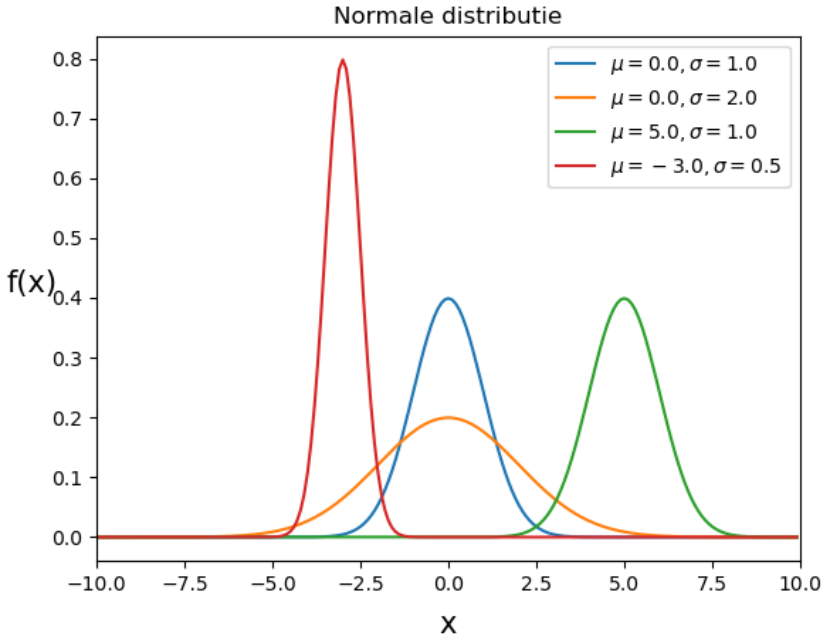
De functie heeft twee parameters, μ en σ, de notering is niet toevallig. De verwachtingswaarde van de normaal verdeling is precies μ en de standaarddeviatie is precies σ.

Over de mathematische beginselen van de Normale verdelingsfunctie gaan we hier verder niet in. Het is wel goed om te weten dat de Normale verdelingsfunctie zonder twijfel de meest belangrijke functie is in de statische data analyse. De verdelingsfunctie komt erg vaak voor. Dat is geen toevalligheid, we zullen later in module 3 zien waarom dit zo is.

Hieronder zie je enkele voorbeelden van de Normale verdeling met verschillende waardes voor μ en σ.

Het is goed om op te merken dat de Normale verdeling een symmetrische continue verdeling is. De meeste stochasten zijn gegroepeerd rond het gemiddelde en hoe meer we van het gemiddelde afwijken, hoe kleiner de kans is dat we een stochast aantreffen.

Voorbeelden van Normaal verdelingen vinden we overal om ons heen. De verdeling van lichaamslengtes van mensen (of bijvoorbeeld olifanten), de grote van zandkorrels op een strand, de luminositeit van bolhopen in het melkwegstelsel.


Normale distributie
$\mu = 0.0, \sigma = 1.0$
$\mu = 0.0, \sigma = 2.0$
$\mu = 5.0, \sigma = 1.0$
$\mu = -3.0, \sigma = 0.5$
f(x)
x
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

Hoofdstuk 3

Voorbeeld opgaves tussentoets I

Lees goed het lijstje door ter voorbereiding voor de tussentoets. **Niet voor alle element op het lijstje zijn oefenopgaves**.

---

1 De leeftijdsverdeling van een groep studenten is:

18.3 19.7 20.4 19.2 18.7 19.4 17.6 20.6 18.5 20.2

a Bereken het gemiddelde, de mediaan, de standaarddeviatie en de variantie.
Antwoord: *Het gemiddelde is 19.3 jaar, de mediaan is 19.3 jaar, de standaarddeviatie is 0.936 jaar, de variantie is 0.876 jaar*^{2}*.*
b De docent van de groep is 44.5 jaar oud.
Bereken nu opnieuw het gemiddelde, de standaarddeviatie en de variantie waarbij je de leeftijd van de docent ook meeneemt.
Antwoord: *Het gemiddelde is 21.6 jaar, de mediaan is 19.4 jaar,*
de standaarddeviatie is 7.31 jaar en de variantie is 53.4 jaar^2*.*
NB Zie je dat het gemiddelde niet veel groter is geworden maar de standaarddeviatie en de variantie wel?

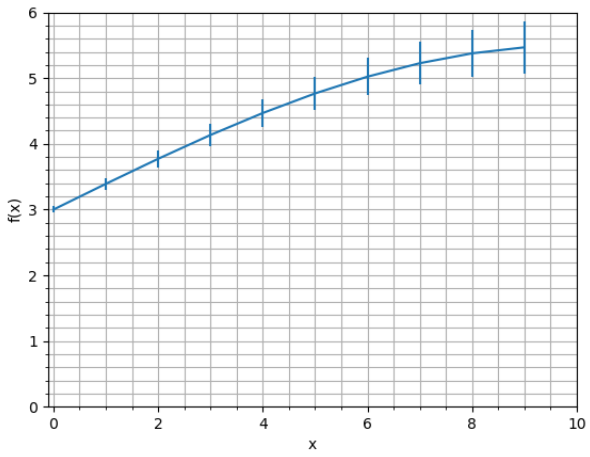
---

2 Lees in de onderstaande grafiek het punt voor x=6 af en noteer het resultaat met de wetenschappelijke notatie.

Antwoord: $f(x) = 5.02 \pm 0.28$ *(Waarschijnlijk lukt het niet om het zo exact af te lezen.)*

---

3 Een bepaald soort knikkers heeft een gemiddelde diameter van 1.4 cm met een variantie van 0.2 cm^2. We willen de gemiddelde diameter bepalen en meten hiervoor de diameter van een enkele knikker op en vinden 1.5 cm. Wat is de fout op deze meting en wat is de onzekerheid?



De fout op de meting is de afstand van de gemeten waarde tot het gemiddelde: 0.1 cm. De onzekerheid is de wortel van de variantie: $\sqrt{0.2}$ cm.
NB Meestal weten we het echte gemiddelde niet en kunnen dan ook de fout niet exact bepalen.*

---

4 We trekken kaarten uit een kaartendek.
a Als we 1 kaart trekken, wat is dan de kans dat we een hartenkaart trekken?
Antwoord: *P(♡) = 1/4*
b Als we 1 kaart trekken, wat is dan de kans dat we een hartenkaart of een schoppenkaart pakken?
Antwoord: *P(♡ of ♠) = P(♡) + P(♠) = 1/2*

---

5 We hebben een zak met gekleurde snoepjes met een tekst erop. Er zijn 6 blauwe snoepjes en 4 rode. Er bestaan drie teksten: “Joepie”, “Hoera!” en “Gefeliciteerd”. De kans verdeling onder de rode snoepjes is P(Joepie) = 0.5 en P(Hoera) = 0.5. Er zijn geen rode snoepjes met gefeliciteerd.
a Als je een snoepje uit de zak pakt, wat is de kans dat je een rood snoepje pakt?
Antwoord: *4/10 = 0.4*
b Als je een snoepje uit de zak pakt, wat is dan de kans dat je een rood snoepje pakt met de tekst Joepie?
Antwoord: *P(rood en joepie) = 0.4 x 0.5 = 0.2*

---

6 Uit een experiment zijn 3 uitkomsten mogelijk voor de gedefinieerde stochast X: {0,1,2}. De kans op uitkomst 0 is: P(0)= 0.56. De kans op uitkomst 2 is P(2) = 0.34.
a Wat is de uitkomstenverzameling van X?
Antwoord: *De uitkomstenverzameling is {0,1,2}*

b Wat is de kans op uitkomst P(1)?
Antwoord: *P(1) = 1 - P(0) - P(2) = 0.10*

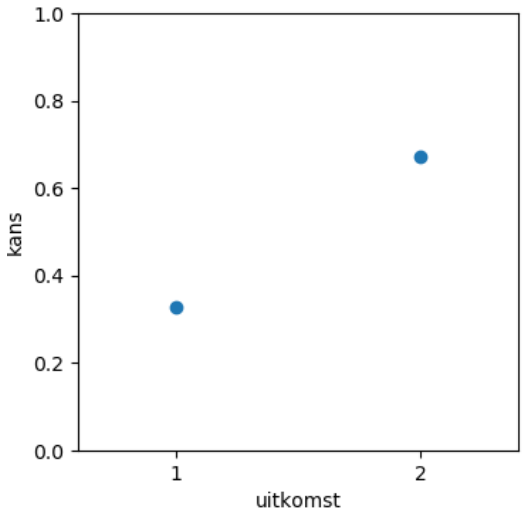
---

7 We definiëren een stochast, x, als de waarde van de worp van een eerlijke dobbelsteen. Wat is de verwachtingswaarde van deze stochast?
Antwoord: $\bar{x} = \frac{1}{6}\sum_{i=1}^{n=6} x_i = \frac{1}{6} \cdot (1+2+3+4+5+6) = \frac{21}{6} = 3.5$.

---

8 We hebben een plankje. Aan een kant van het plankje zetten we een 1, aan de andere kant een 2. We laten het plankje 1000 keer vallen van steeds dezelfde hoogte en we houden bij welke kant boven ligt als het plankje gevallen is. Van de 1000 keer ligt nummer 1 slechts 327 keer boven. Geef de kans verdeling van de twee uitkomsten grafisch weer.

Antwoord: *We gebruiken de Frequentist kans: P(1) = 327/1000 = 0.327 en de*



complement-regel dat P(2)= 1- P(1) = 0.673.

---

9 Je gooit 10 keer met een dobbelsteen.
a Wat is de kans dat je precies 6 keer een 6 gooit?
Antwoord: *Gebruik de binomiaal vergelijking:*

$$P(k=6; n=10, p=1/6) = \frac{10!}{6!(4)!}(1/6)^6(5/6)^4 = 0.0022 \tag{3.1}$$

b Wat is de kans dat je precies 1 keer een 6 gooit?
Antwoord: *P(k=1;n=10,p=1/6) = 0.32*
c Wat is de kans dat je minder dan 3 keer een 6 gooit?
Antwoord: *P(<3maal een zes) = P(0;10,1/6) + P(1;10,1/6) + P(2;10,1/6) = 0.162 +*

0.323 + 0.291 = 0.78
d Wat is de kans dat vaker dan 2 keer een 6 gooit?
Antwoord: *P(>2maal een zes) = 1-0.78 = 0.22*

---

10 Een raketschild houdt 99% van de raketten tegen.
a Als door de tegenstander 20 raketten worden afgevuurd, wat is dan de kans dat het raketschild alle 20 tegen houdt?
Antwoord: $P = 0.99^{20} = 0.82$
b Als er 50 raketten worden afgevuurd door de tegenstander, hoeveel raketten worden er dan gemiddeld tegengehouden?
Antwoord: $E = 50 \cdot 0.99 = 49.5$

---

11 In een call-center komen gemiddeld 100 telefoontjes per dag.

telefoontjes per dag?
Antwoord: *Gebruik de Poisson verdeling:* $\sigma = \sqrt{100} = 10$.
b Als er op een dag 70 telefoontjes binnenkomen. Is dat gek?
Antwoord: *100-70/10 = 3* σ *Het is zeker uitzonderlijk het ligt* 3σ *van het gemiddelde af.*

---

12 In een stad gebeuren jaarlijks 1020 ongelukken. Het afgelopen jaar zijn er maar 900 ongelukken geweest. De autoriteiten claimen dat dit komt door nieuwe regels in het verkeer.
a Denk je dat deze verklaring klopt?
Antwoord: *De spreiding op het aantal ongelukken is* $\sqrt{1020} = 42$. *Het verschil is uitgedrukt in standaarddeviaties:* $\frac{1020-900}{42}\sigma = 3.8\sigma$. *Dat is zeker uitzonderlijk, maar het komt natuurlijk weleens voor. Het is goed om voorzichtig te zijn met zo' n uitspraak.*
b Als de getallen tien keer zo klein zouden zijn (102 en 90), zou je denken dat de verklaring dan nog klopt?
Antwoord: *De standaarddeviatie is nu* $\sigma = 10$. *Het verschil is dus net iets groter dan* 1σ, *zo'n afwijking komt vaak voor. De uitspraak lijkt ongegrond.*

---

13 We verwachten op een dag gemiddeld 4.3 poststukken bij een klein bedrijf.
a Op een dag komen er wel 7 binnen. Reken de kans uit dat dit gebeurt.
Antwoord: $P(k = 7; \lambda = 4.3) = 0.073$
b Reken de kans uit dat er 0 binnenkomen.
Antwoord: $P(k = 0; \lambda = 4.3) = 0.014$