Statistiek MBW/KW deel 1 week 5

- Negatief exponentiële verdeling
- Continuïteitscorrectie
- Wet van de grote aantallen

Poissonverdeling Definitie

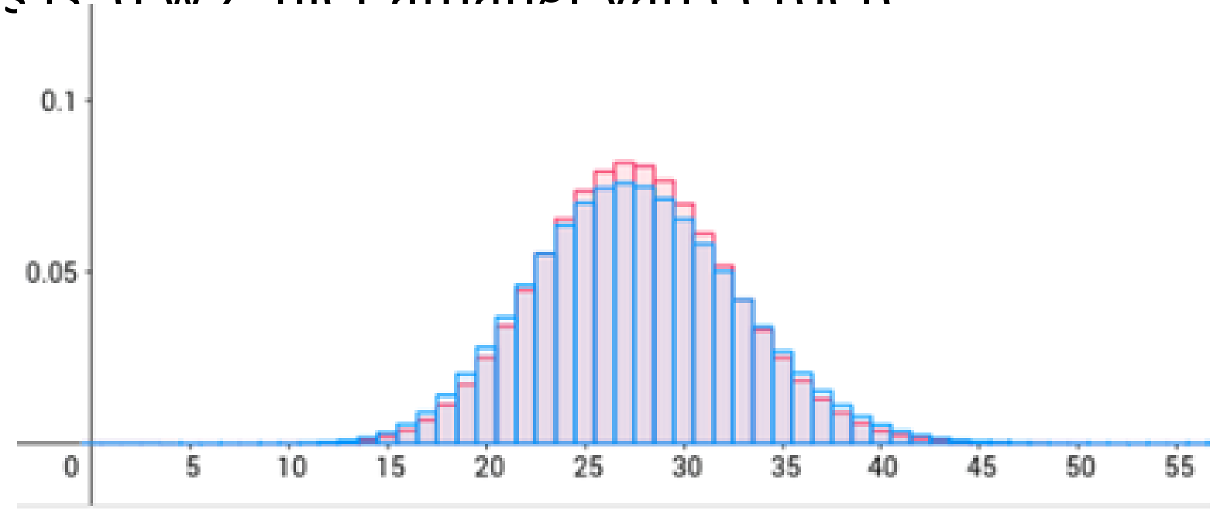
Poissonverdeling (1838). Voor een verschijnsel dat gemiddeld μ keer per (tijd- of plaats)eenheid plaatsvindt is de kans op k successen in één eenheid:

$$P(\underline{k} = k) = \frac{\mu^k}{k!} e^{-\mu}.$$

Voorwaarde is dat het verschijnen geheugenloos is, d.w.z. niet afhangt van eerdere optredens.

De Poissonverdeling Poisson(μ) kun je ook zien als een benadering van binom($n; p$), voor grote $n \to \infty$, en kleine $p \to 0$, waarbij de gemiddelde waarde $\mu = np$ vast is.

Dit is een goede benadering als
$$\boldsymbol{n \geq 20, \mu = np < 5 \text{ en } p \leq 0,05.}$$



Binom($n = 197$; $p = 0{,}14$)
Poisson($\mu = 197 \times 0{,}14 = 27{,}58$)

Binompdf(197 , 0.14 , 30) = 0,06992
Poissonpdf(27.58 , 30) = 0,06551

Negatief exponentiële verdeling

De Poissonverdeling geeft je de kans op het **aantal gebeurtenissen** in een bepaalde tijd, maar stel dat je geïnteresseerd bent in de **tijd tussen twee gebeurtenissen**.

Als op tijdstip $t = 0$ een gebeurtenis plaatsvindt, wat is dan de kans dat het minstens t tijdseenheden duurt tot de volgende gebeurtenis: $P\left(\underline{t} > t\right)$?

Dat is de kans dat er tussen de tijdstippen $t = 0$ en t géén gebeurtenis plaatsvindt. Neem in de Poissonverdeling als tijdseenheid [0,t] en het verwachte aantal gebeurtenissen in die tijdseenheid: $\tilde{\mu} = \mu t$, dan is $P\left(\underline{k} = 0\right) = P\left(\underline{t} > t\right) = \frac{(\mu t)^0}{0!} e^{-\mu t} = e^{-\mu t}$.

De kansverdelingsfunctie (cdf) is de kans dat de eerstvolgende gebeurtenis wél binnen 0 en t optreedt: $F(t) = P\left(\underline{t} < t\right) = 1 - P\left(\underline{t} > t\right) = 1 - e^{-\mu t}$.
De kansdichtheidsfunctie (pdf) is de afgeleide daarvan: $f(t) = F'(t) = \mu e^{-\mu t}$, maar die wordt in de praktijk niet gebruikt, net als bij de normale verdeling.

Negatief exponentiële verdeling - Eigenschappen

Deze **negatief exponentiële verdeling** (a.k.a. **exponentiële verdeling**) is een continue verdeling met één parameter, die gerelateerd is aan de Poissonverdeling.

De formules voor kansverdeling en kansdichtheid gelden voor $t \geq 0$, voor $t < 0$ zijn ze 0.

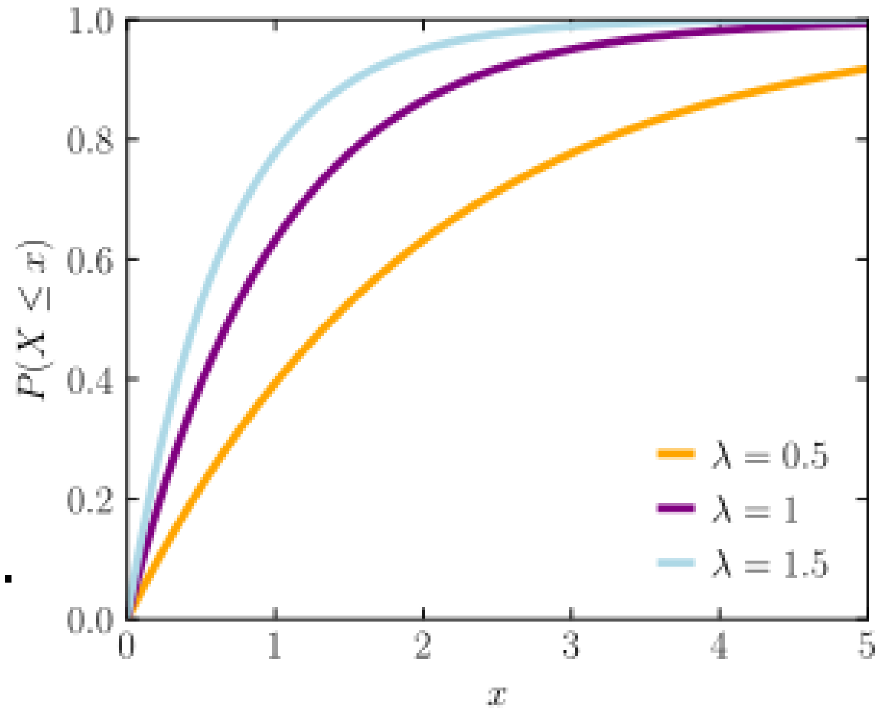
De notatie voor de negatief exponentiële verdeling is: $\underline{t} \sim \exp(\mu)$.

De verwachtingswaarde is $E(\underline{t}) = \frac{1}{\mu}$ en

de standaarddeviatie is $\sigma(\underline{t}) = \frac{1}{\mu}$.

De exponentiële verdeling zit niet op je rekenmachine, berekeningen doe je met de kansverdelingsfunctie (cdf)

$$\boldsymbol{F(t) = P(\underline{t} < t) = 1 - e^{-\mu t}}.$$



Negatief exponentiële verdeling

Voorbeeld: Op buslijn 17 rijden drie bussen per uur. Door stoplichten en verschillende stoptijden rijden de bussen niet precies elke 20 minuten, maar er kan worden aangenomen dat er gemiddeld 3 bussen per uur langskomen.
De tijdseenheid is één uur en $\mu = 3$. De kans dat de eerstvolgende bus binnen t uur komt is (verdelingsfunctie van de exponentiële verdeling)

$$F(t) = P(\underline{t} < t) = 1 - e^{-3t}.$$

De kans op hoogstens tien minuten wachten is $P\left(\underline{t} < \frac{10}{60}\right) = 1 - e^{-3\cdot\frac{1}{6}} = 0{,}3935.$

De kans op minstens een half uur wachten is $P\left(\underline{t} > \frac{1}{2}\right) = e^{-\frac{3}{2}} = 0{,}2231.$

De kans dat de eerstvolgende bus tussen 10 en 15 minuten na nu komt is
$P\left(\frac{10}{60} < \underline{t} < \frac{15}{60}\right) = F\left(\frac{15}{60}\right) - F\left(\frac{10}{60}\right) = \left(1 - e^{-3\cdot\frac{15}{60}}\right) - \left(1 - e^{-3\cdot\frac{10}{60}}\right) = e^{-\frac{1}{2}} - e^{-\frac{3}{4}} =$
$0{,}1342.$

Negatief exponentiële verdeling

Voorbeeld: Op buslijn 17 rijden drie bussen per uur. Door stoplichten en verschillende stoptijden rijden de bussen niet precies elke 20 minuten, maar er kan worden aangenomen dat er gemiddeld 3 bussen per uur langskomen.

De kansen kun je ook uitrekenen met de Poissonverdeling:

De kans op hoogstens tien minuten wachten is (tijdseenheid 10 minuten, $\mu = 3 \cdot \frac{10}{60}$)
$P(\underline{k} \geq 1) = 1 - \text{poissonpdf}(\mu = 0.5\,; k = 0) = 0{,}3935.$

De kans op minstens een half uur wachten is (tijdseenheid 0,5 uur, $\mu = 3 \cdot 0{,}5$)
$P(\underline{k} = 0) = \text{poissonpdf}(\mu = 1.5\,; k = 0) = 0{,}2231.$

De kans dat de eerstvolgende bus tussen 10 en 15 minuten na nu komt is $P(\underline{k} \geq 1) - P(\underline{k} \geq 1)$, waarbij de eerste kans met tijd 15 minuten en de tweede met 10 minuten is genomen:

$$P(\underline{k} \geq 1) - P(\underline{k} \geq 1) = \left(1 - \text{poissonpdf}(\mu = 0.75\,; k = 0)\right) - \left(1 - \text{poissonpdf}(\mu = 0.5\,; k = 0)\right) = 0{,}1342.$$

Negatief exponentiële verdeling Tentamenopgave

Voorbeeld 4 (oude tentamenopgave) Op een kleine luchthaven is slechts één securitydoorgang. Het aantal passagiers dat zich per minuut bij de security meldt, kan worden beschreven m.b.v. een Poissonverdeling met μ = 4 (passagiers per minuut).
De securitydoorgang heeft een capaciteit met grootte g, d.w.z. als zich per minuut meer passagiers melden, dan ontstaan er problemen bij de afhandeling van passagiers en hun bagage, de waarde van g is niet bekend.
4a. Hoe groot is de kans dat zich gedurende één minuut meer dan 5 passagiers melden?
4b. Hoe groot is de kans dat zich gedurende een uur meer dan 250 passagiers melden?
4c. De grootte van de capaciteit g kan worden bepaald op basis van de eis dat de kans op problemen bij de afhandeling van passagiers en bagage hoogstens gelijk mag zijn aan 0,01. Hoe groot is de capaciteit g?
4d. Hoe groot is de kans dat het tijdsinterval tussen twee bij de security arriverende passagiers meer dan 15 seconden, maar minder dan 30 seconden bedraagt? Maak gebruik van de exponentiële verdeling.

Negatief exponentiële verdeling Tentamenopgave

4a. Hoe groot is de kans dat zich gedurende een minuut meer dan 5 passagiers melden?

Het gaat hier om aantallen $\underline{k}$, dus gebruik hiervoor de Poissonverdeling met $\mu = 4$ passagiers per minuut, met één minuut als tijdeenheid:

$$P(\underline{k} > 5) = 1 - P(\underline{k} \leq 5) = 1 - \text{poissoncdf}(\mu = 4\,, k = 5) = 0{,}21486\ldots$$

4b. Hoe groot is de kans dat zich gedurende een uur meer dan 250 passagiers melden?

Neem een uur als tijdseenheid, dan is het gemiddeld aantal passagiers in een uur $\mu = 60 \cdot 4 = 240$, dus de kans dat er in één uur meer dan 250 passagiers komen is volgens de Poissonverdeling

$$P\left(\underline{k} > 250\right) = 1 - P\left(\underline{k} \leq 250\right) = 1 - \text{poissoncdf}(240\,, 250) = 0{,}24711\ldots$$

Negatief exponentiële verdeling Tentamenopgave

4c. De grootte van de capaciteit g kan worden bepaald op basis van de eis dat de kans op problemen bij de afhandeling van passagiers en bagage hoogstens gelijk mag zijn aan 0,01. Hoe groot is de capaciteit g?

De capaciteit g is het (kleinste!) getal waarvoor geldt dat de kans op overschrijden van de capaciteit hoogstens 0,01 mag zijn: $P(\underline{k} > g) \leq 0{,}01$, ofwel
$1 - P(\underline{k} \leq g) \leq 0{,}01$, ofwel $P(\underline{k} \leq g) > 0{,}99$.
Met tabel C5 uit Buijs vind je dan dat $g = 9$.

Of door uitproberen met de rekenmachine (maak een tabel):
Poissoncdf(4 , 9) = 0,99186
Poissoncdf(4 , 8) = 0,97864

Of los op met Solver: Poissoncdf(4 , n) = 0,99 en controleer: Poissoncdf(4 , 9) = 0,9918

Negatief exponentiële verdeling Tentamenopgave

4d. Hoe groot is de kans dat het tijdsinterval tussen twee bij de security arriverende passagiers meer dan 15 seconden, maar minder dan 30 seconden bedraagt? Maak gebruik van de *exponentiële verdeling*.

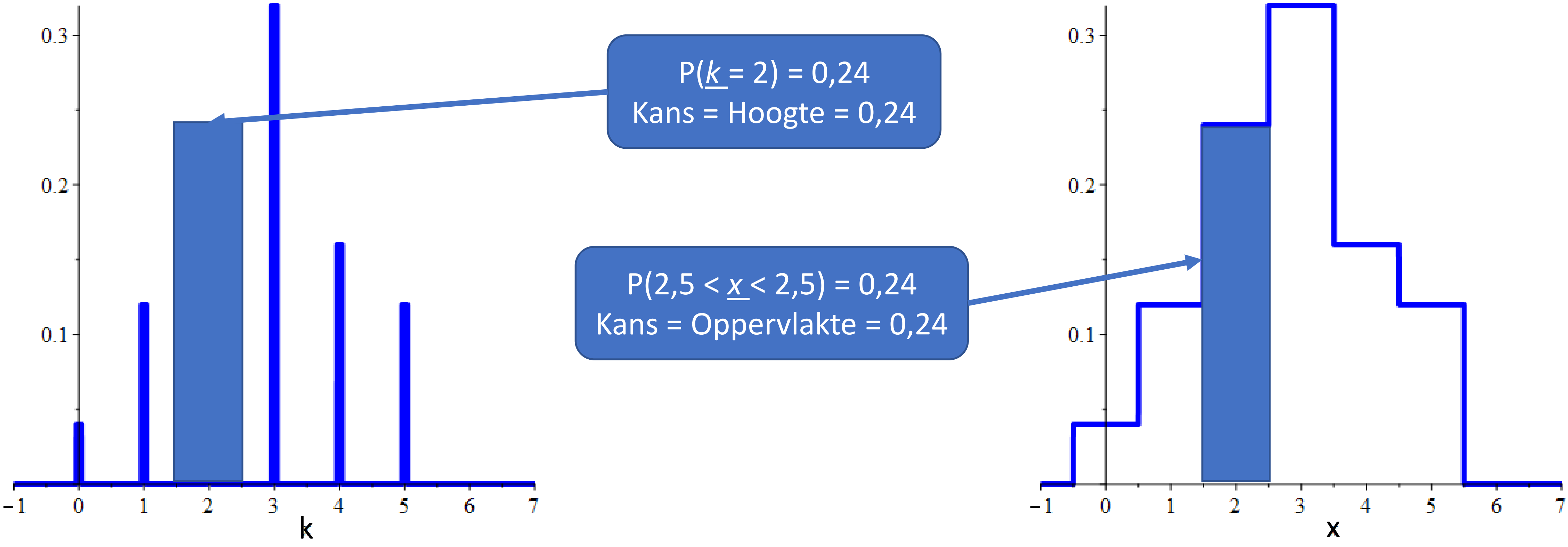
De tijd tussen twee passagiers wordt beschreven door exp(4), waarbij de tijd in minuten is. De gevraagde kans is

$$P\left(\frac{15}{60} = 0{,}25 < \underline{t} < \frac{30}{60} = 0{,}50\right) = P(\underline{t} < 0{,}50) - P(\underline{t} < 0{,}25)$$

$$= F(0{,}50) - F(0{,}25)$$
$$= (1 - e^{-4\cdot 0{,}50}) - (1 - e^{-4\cdot 0{,}25}) = e^{-1} - e^{-2} = 0{,}23254.$$

Discrete verdeling benaderen door een continue verdeling. Continuïteitscorrectie

Soms is het verstandig om een discrete kansverdeling te benaderen door een continue.
Reden: Voor continue verdelingen zijn meer technieken (differentiëren/integreren).
Maak van een discrete variabele $\underline{k}$ met kansfunctie een continue variabele $\underline{x}$ met kansdichtheidsfunctie door staafjes te vervangen door balkjes met breedte 1.



$P(1 \leq \underline{k} \leq 3) = 0{,}68$

Continuïteitscorrectie: $P(1 - 0{,}5 \leq \underline{x} \leq 3 + 0{,}5) = 0{,}68$

Centrale limietstelling (Wet van de grote aantallen)

Trek n getallen $\underline{x}_1, \ldots, \underline{x}_n$ uit een vaste verdeling waarvan gemiddelde μ en standaard-deviatie σ gegeven zijn (verdere details van de verdeling zijn niet van belang), dan voldoen het gemiddelde $\underline{x}_{gem}$ en de som $\underline{x}_{som}$ van deze getallen goed aan een normale verdeling, als n groot genoeg is (richtlijn: $n \geq 20$). Er geldt:

$$\underline{x}_{gem} = \underline{\bar{x}} \sim N\left(\mu\,;\,\frac{\sigma}{\sqrt{n}}\right)$$

$$\underline{x}_{som} \sim N(n\mu\,;\sqrt{n}\sigma)$$

De tweede regel volgt uit de eerste, want $\underline{x}_{som} = n\underline{x}_{gem}$

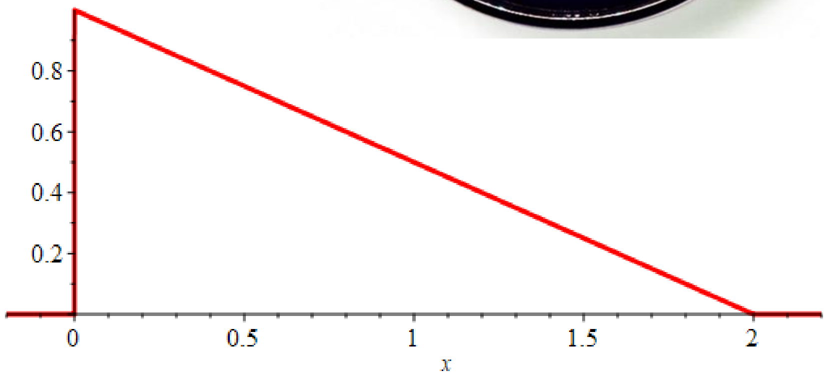
Als de verdeling waaruit wordt getrokken toevallig zelf een normale verdeling is dan geldt de benadering zelfs exact.

De centrale limietstelling kun je gebruiken om het gemiddelde of som van herhaalde trekkingen uit telkens dezelfde kansverdeling te benaderen met een normale verdeling

Voorbeeld: Schieten op een schietschijf (Reader p.10)

Je schiet op een schietschijf. De afstand van het inslagpunt tot het midden is $\underline{x}$. De kansdichtheidsfunctie is

$$f(x) = \begin{cases} 1 - \dfrac{x}{2} & \text{als } 0 \leq x \leq 2 \\ 0 & \text{elders} \end{cases}$$



(Dit is een aanname, niet noodzakelijk realistisch)

Je schiet "raak" als $\underline{x} \leq \frac{1}{2}$. De kans hierop is

Met de GR: $P\left(\underline{x} \leq \frac{1}{2}\right) = \int_0^{1/2} \left(1 - \frac{x}{2}\right) dx = \mathrm{fnInt}(1 - x/2, x, 0{,}0.5) = 0{,}4375.$

Of met Wiskunde B: $P\left(\underline{x} \leq \frac{1}{2}\right) = \int_0^{1/2} \left(1 - \frac{x}{2}\right) dx = \left[x - \frac{1}{4}x^2\right]_0^{1/2} = \frac{1}{2} - \frac{1}{4}\left(\frac{1}{2}\right)^2 = \frac{7}{16} =$ $0{,}4375.$

Voorbeeld: Schieten op een schietschijf

De verwachtingswaarde, de variantie en de standaarddeviatie zijn met de GR:

$$\mu = \mathrm{E}(\underline{x}) = \int_0^2 \boldsymbol{x}\left(1 - \frac{1}{2}x\right)dx = \mathrm{fnInt}\left(x * \left(1 - \frac{x}{2}\right), x, 0, 2\right) = 0{,}666666$$

$$\mathrm{Var}(\underline{x}) = \mathrm{E}\left((\underline{x} - \mu)^2\right) = \mathrm{fnInt}((x - \mathbf{0.66666})\text{^}2 * (1 - x/2), x, 0{,}2) = 0{,}222222$$

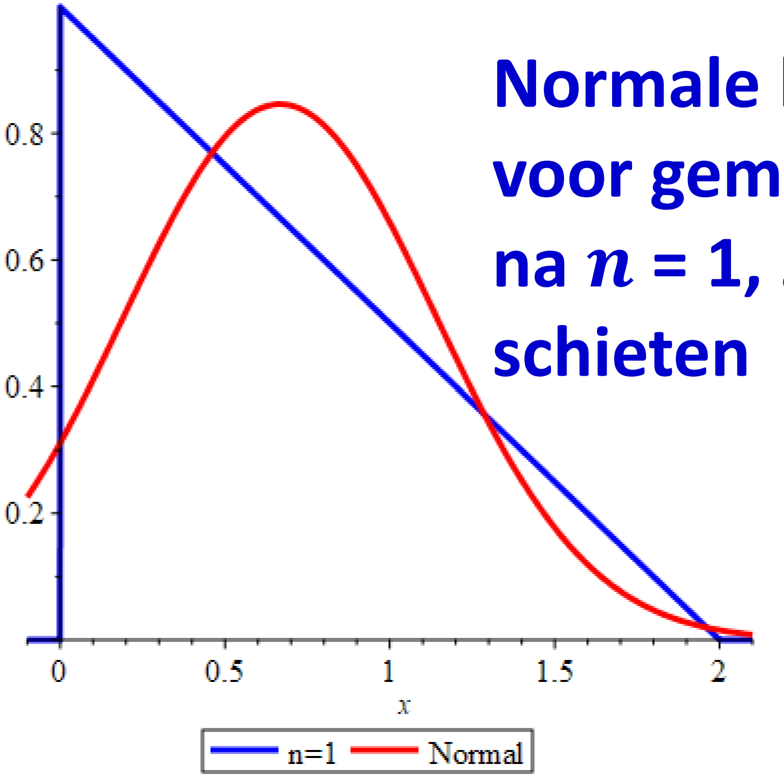
$$\sigma = \sqrt{0{,}22222} = 0{,}4714.$$

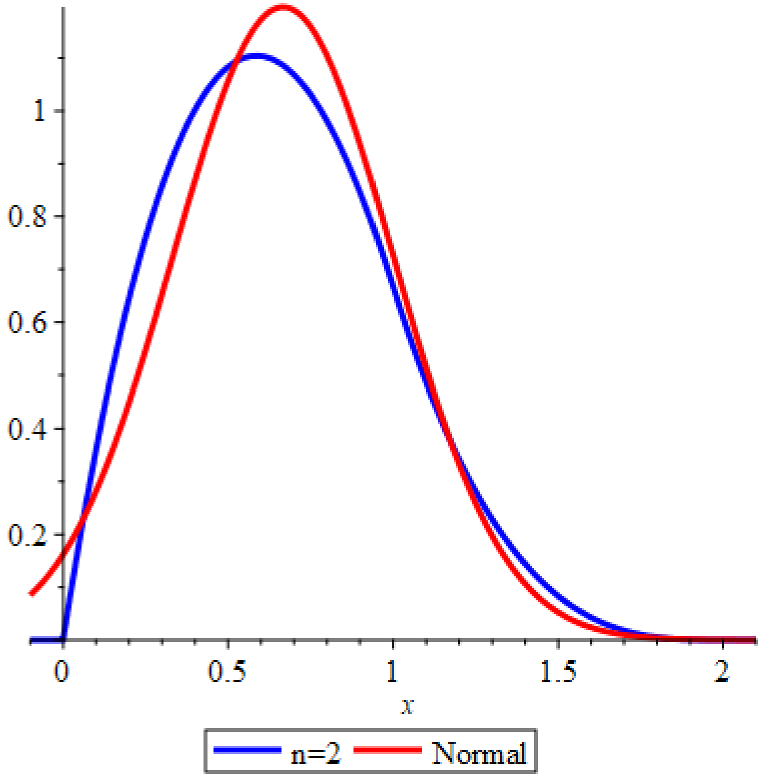
Of met Wiskunde B:

$$\mu = \mathrm{E}(\underline{x}) = \int_0^2 \boldsymbol{x}\left(1 - \frac{1}{2}x\right)dx = \left[\frac{1}{2}x^2 - \frac{1}{6}x^3\right]_0^2 = 2 - \frac{8}{6} = \mathbf{\frac{2}{3}} = 0{,}6667,$$

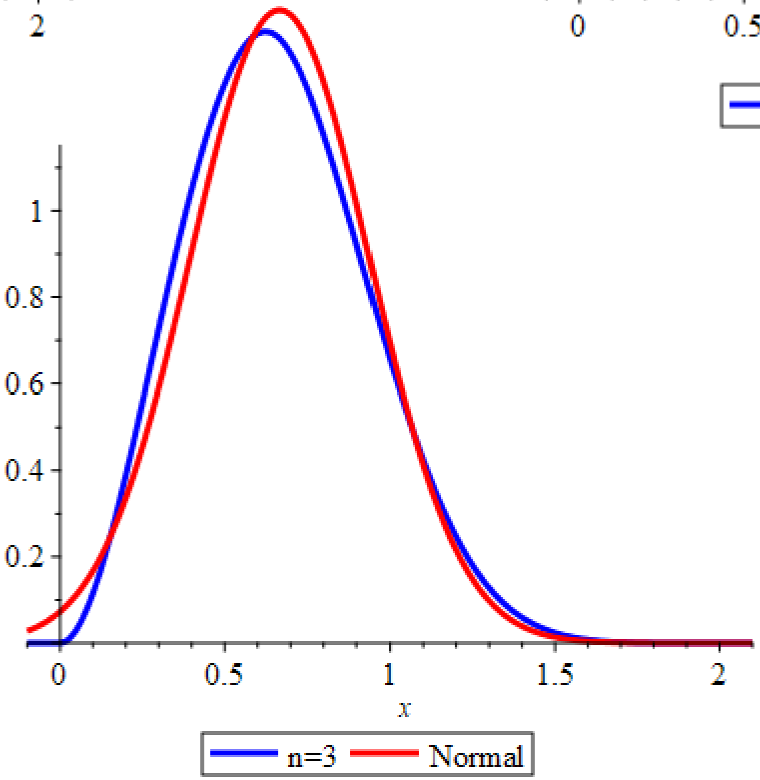
$$\mathrm{Var}(\underline{x}) = \mathrm{E}(\underline{x}^2) - \mu^2 = \int_0^2 x^2\left(1 - \frac{1}{2}x\right)dx - \frac{4}{9} = \left[\frac{1}{3}x^3 - \frac{1}{8}x^4\right]_0^2 - \frac{4}{9} = \mathbf{\frac{2}{9}} = 0{,}2222,$$

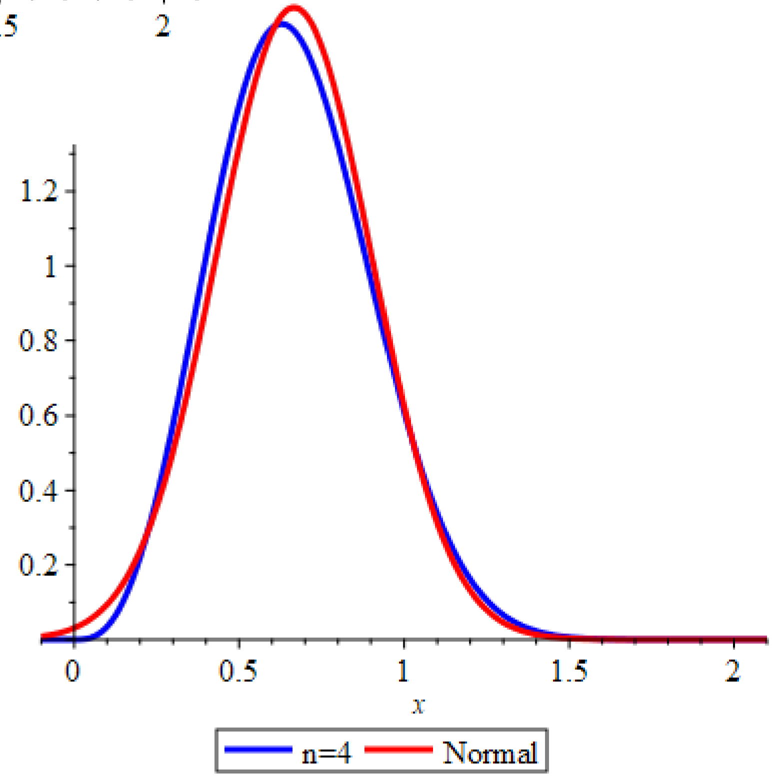
Normale benadering voor gemiddelde score na n = 1, 2, 3, 4 keer schieten





In blauw staan de exacte kansdichtheid-functies die lastig zijn uit te rekenen, in rood de benaderingen met normale verdelingen (die ook met kleine kans negatieve afstanden toelaten, wat eigenlijk niet kan).





Voorbeeld: Schieten op een schietschijf

De kansdichtheidsfunctie van de gemiddelde score (raak = 1, mis = 0) over een aantal keren schieten is niet exact te bepalen, maar uit de centrale limietstelling volgt wel dat de gemiddelde waarde na n keer schieten zich gedraagt als $N\left(\mu\ ;\ \frac{\sigma}{\sqrt{n}}\right) = N\left(\frac{2}{3}\ ;\ \sqrt{\frac{2}{9n}}\right)$.

De kans dat na 10 keer schieten de gemiddelde score kleiner dan 3/4 is, is bijvoorbeeld

$$P\left(\underline{x}_{gem} < \frac{3}{4}\right) \approx \text{Normalcdf}\left(-10^{10}, \frac{3}{4}, \frac{2}{3}, \sqrt{\frac{2}{9 \cdot 4}}\right) = 0{,}6382.$$

De exacte waarde is $\frac{83833}{129024} = 0{,}6497$. De benadering is al redelijk, ondanks de kleine n.

Hoe vaak moet je schieten zodat deze kans groter is dan 0,90?

$$z = \text{invNorm}(0.90) = 1{,}28155$$

$$1{,}28155 = z = \frac{x-\mu}{\sigma} = \left(\frac{3}{4} - \frac{2}{3}\right) \Big/ \sqrt{\frac{2}{9n}} \quad \rightarrow \quad n \geq 52{,}56 \quad \rightarrow n = 53$$

Of met de Solver: $\text{Normalcdf}\left(-10^{10}, \frac{3}{4}, \frac{2}{3}, \sqrt{\frac{2}{9 \cdot n}}\right) = 0{,}90 \rightarrow \quad n \geq 52{,}56 \quad \rightarrow n = 53$