Tentamen Statistiek MBW/KW deel 1, tweede kans
Tentamen Statistiek MBW/KW deel 1, vierde kans (2023)

Afdeling: Propedeuse KW/MBW 2022-2023 en 2023-2024
Examinator: Dr. ir. D.A.M.P. Blom
Datum: 18 oktober 2024, **duur tentamen: 3 uur**

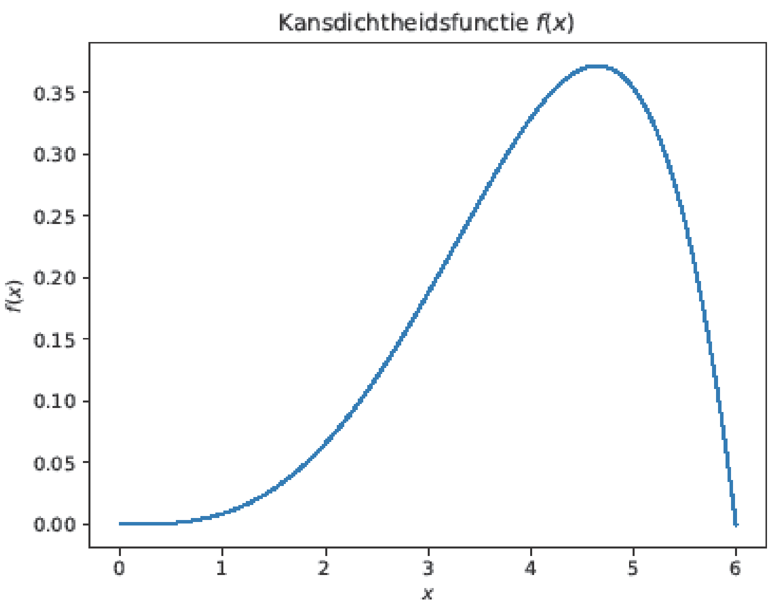
1. **Alle antwoorden moeten gemotiveerd worden**!
2. Rond eindantwoorden (kommagetallen) af op *vier* decimalen, tenzij anders vermeld.
3. Boeken, reader en aantekeningen mogen worden geraadpleegd.
4. De aanwezigheid van *communicatieapparatuur* is niet toegestaan.
5. Het gebruik van een (grafische) rekenmachine met statistische programmatuur is toegestaan. Het *statistische* gebruik van deze rekenmachine is mogelijk ingeperkt. Let op de aanwijzingen!
6. **De opgaven dienen na afloop van het tentamen ingeleverd te worden.**

Dit tentamen bestaat uit vier opgaven (30, 25, 25, 20 punten). Score = Puntentotaal/10

Opgave 1 (Totaal 30 punten)

De kansvariabele $\underline{x}$ heeft de volgende kansdichtheidsfunctie:

$$f(x) = \begin{cases} \frac{1}{3888}x^3(36-x^2), & \text{als } 0 \le x \le 6 \\ 0\,, & \text{elders} \end{cases}$$



1a. [4pt] Leg uit waarom de functie $f(x)$ een goed gedefinieerde kansdichtheidsfunctie is. Maak hierbij gebruik van een berekening.

Om te laten zien dat $f(x)$ goed gedefinieerd is, moeten we twee voorwaarden checken:

1. $f(x) \geq 0$: hier is duidelijk aan voldaan in het geval van $x < 0$ en $x > 6$ (1pt). Merk tevens op dat voor $0 \leq x \leq 6$ geldt dat $x^3 \geq 0$ en $36 - x^2 \geq 0$. De term $\frac{1}{3888}x^3(36 - x^2)$ is dus het product van niet-negatieve termen, dus ≥ 0.
2. $\int_{-\infty}^{\infty} f(x)dx = \int_0^6 \frac{1}{3888}x^3(36 - x^2)dx = fnInt(\frac{1}{3888}x^3(36 - x^2), x, 0{,}6) = 1$
(1pt voor realisatie integraal van $-\infty$ to ∞, 1pt voor verdere correcte berekening.)

1b. [8pt] Bereken de verwachtingswaarde en de standaarddeviatie van de kansvariabele $\underline{x}$.

- **Verwachtingswaarde:**
$E[\underline{x}] = \int_{-\infty}^{\infty} xf(x)dx = \int_0^6 x * \frac{1}{3888}x^3(36 - x^2)dx = \int_0^6 \frac{1}{3888}x^4(36 - x^2)dx = fnInt(\frac{1}{3888}x^4(36 - x^2), x, 0, 6) \approx 4.1143$ (3pt)
- **Standaarddeviatie:**
$Var(\underline{x}) = \int_{-\infty}^{\infty}(x - E[\underline{x}])^2 f(x)\,dx = \int_0^6 (x - 4.1143)^2 * \frac{1}{3888}x^3(36 - x^2)\,dx \approx$ 1.07265 (3pt)
$\Rightarrow \sigma_{\underline{x}} = \sqrt{Var(\underline{x})} = \sqrt{1.0727} \approx 1.0357$ (2pt)

1c. [4pt] Bereken de kans dat de waarde van $\underline{x}$ groter dan 3 is.

Omdat $\underline{x}$ een continue kansvariabele is, geldt dat $P(\underline{x} > 3) = P(\underline{x} \geq 3)$. (1pt)

We berekenen deze kans als volgt: $P(\underline{x} > 3) = P(\underline{x} \geq 3) = \int_3^6 f(x)dx = \int_3^6 \frac{1}{3888}x^3(36 - x^2)dx \approx 0.8438$ (2pt voor kans zien als integraal, 1pt voor complete uitwerking)

1d. [6pt] Stel we bekijken vijftig onafhankelijke trekkingen van de kansvariabele $\underline{x}$. Hoe groot is de kans dat minstens 42 van deze trekkingen een waarde groter dan 3 heeft? Indien je geen antwoord hebt bij 1c, mag je aannemen dat de kans gelijk is aan 0.85 dat een enkele trekking groter dan 3 is.

Voor elke trekking van $\underline{x}$ kunnen we een waarde groter dan 3 definieren als succes (1) en een waarde kleiner dan of gelijk aan 3 als mislukking (0). (1pt)
De succeskans voor een individuele trekking is gelijk aan 0.8438 (of 0.85) (1pt)
Omdat we hier kijken naar onafhankelijke trekkingen van een kansvariabele, kunnen we de kansvariabele $\underline{y}$ beschouwen die het aantal successen telt (1pt)
Deze kansvariabele $\underline{y}$ is binomiaal verdeeld ($n = 50$ en $p = 0.8438\ (of\ 0.85)$) (1pt)
De kans op minstens 42 successen is dan $P(\underline{y} \geq 42) = 1 - P(\underline{y} \leq 41) = 1 - binomcdf(50, 0.8438, 41) \approx 1 - 0.3783 = 0.6217$ (2pt)

1e. [8pt] Tenslotte wordt er gekeken naar het gemiddelde van de vijftig onafhankelijke trekkingen uit 1d. Gebruik de centrale limietstelling om de kans uit te rekenen dat deze gemiddelde waarde tussen 4,0 en 4,2 zit.
De centrale limietstelling zegt dat het gemiddelde $\overline{x}$ van de vijftig onafhankelijke trekkingen uit dezelfde kansverdeling bij benadering normaal verdeeld is met verwachtingswaarde $E[\underline{x}] = 4.1143$ en standaarddeviatie $\frac{\sigma_x}{\sqrt{n}} = \frac{\sigma_x}{\sqrt{50}} \approx 0.1465$.
(2pt voor normale verdeling, 2pt voor verwachtingswaarde, 2pt voor standaarddeviatie).

De kans dat deze gemiddelde waarde tussen 4.0 en 4.2 ligt is daarom gelijk aan $P(4.0 \leq \overline{x} \leq 4.2) = normalcdf(4.0; 4.2;\ 4.1143;\ 0.1465) \approx 0.5031.$ (2pt)

Opgave 2 (Totaal 25 punten)

Binnen de Divisie Personeel en Organisatie Defensie (DPOD) houdt de afdeling TOS (Trends, Onderzoek en Statistiek) zich bezig met sociaalwetenschappelijk onderzoek onder Defensiepersoneel. Een van de belangrijke thema's binnen Defensie is het werken aan een schaalbare krijgsmacht, bijvoorbeeld door het werven van meer reservisten.

We nemen aan dat aanmeldingen van reservisten binnenkomen volgens een Poissonproces.

2a [5pt] Gedurende de eerste acht maanden van 2024 (244 dagen) hebben zich 1.087 mensen aangemeld om reservist te worden. Bereken de kans dat er op een willekeurige dag minstens 3 mensen zich aanmelden als reservist.

Laat λ het gemiddeld aantal aanmeldingen per tijdseenheid (een dag) zijn, en t het aantal tijdseenheden (aantal dagen). In dit geval dus $\lambda = \frac{1.087}{244} \approx 4{,}455$ en $t = 1.$ (2pt)

Stel $\underline{k}$ is het aantal aanmeldingen op een willekeurige dag, dan willen we de kans $P(\underline{k} \geq 3)$ berekenen: (1pt)

$P(\underline{k} \geq 3) = 1 - P(\underline{k} \leq 2) = 1 - poissoncdf(\lambda t; 2) = 1 - 0.1787 \approx 0.8213$ (2pt)

2b [4pt] Bereken met behulp van de exponentiële verdeling de kans dat op een willekeurige dag er geen nieuwe aanmeldingen binnenkomen.

Laat $\underline{T}$ de tijd zijn tot de volgende aanmelding. In dat geval geldt als er gedurende een dag geen nieuwe aanmeldingen binnenkomen, dit ook kan worden weergeven met $T > 1.$ (1pt)

Oftewel: $P(\underline{x} = 0) = P(T > 1) = e^{-\lambda t} = e^{-4{,}455} \approx 0.012$ (2pt)

<u>Alternatief</u>: correcte berekening met Poissonformule (3pt)

2c [3pt] Op een willekeurige dag komen er vijf aanmeldingen binnen. Bereken de kans hierop.

Laat $\underline{x}$ is het aantal aanmeldingen op een willekeurige dag, we willen nu de kans $P(\underline{x} = 5)$ berekenen. (1pt)

$P(\underline{x} = 5) = e^{-\lambda t} * \frac{(\lambda t)^5}{5!} = 0.170$ (2pt)

2d [2pt] Op een willekeurige dag komen er vijf aanmeldingen binnen. Leg zonder berekening uit wat de kans hierop is, gegeven dat dit de vorige dag ook is gebeurd.

De aankomsten in een Poissonproces vinden onafhankelijk van elkaar plaats, en de tijd tot de volgende aankomst is onafhankelijk van wat er in het verleden heeft plaatsgevonden (dit is de 'geheugenloosheid' van de exponentiele verdeling). (1pt)

De kans hierop is gelijk aan het antwoord op vraag 2c. (1pt)

2e [5pt] Is de kans op tien aanmeldingen in twee dagen groter dan / gelijk aan / kleiner dan de kans op vijf aanmeldingen op de eerste dag en vijf aanmeldingen op de tweede dag? Beargumenteer je antwoord met een berekening.

De kans op tien aanmeldingen in twee dagen is groter dan de kans op vijf aanmeldingen op zowel de eerste als de tweede dag. (1pt)

Dit heeft te maken met het feit dat de tien aanmeldingen op meerdere manieren kunnen zijn verdeeld over beide dagen (1pt)

Stel $\underline{x}$ het aantal aanmeldingen in twee dagen. De kans op tien aanmeldingen in twee dagen is dan:

$\lambda = \frac{1087}{244},\ t = 2, k = 10: P(\underline{k} = 10) = e^{-\lambda t} * \frac{(\lambda t)^{10}}{10!} = 0.117$ (1pt)

Stel $\underline{k}_1$ en $\underline{k}_2$ het aantal aanmeldingen op dag 1 en dag 2. De kans op vijf aanmeldingen op zowel dag 1 als dag 2 is dan:

$$P(\underline{k}_1 = 5, \underline{k}_2 = 5) = P(\underline{k}_1 = 5)P(\underline{k}_2 = 5) \approx 0.170 * 0.170 = 0.0289$$

Dit laat zien dat de kans op tien aanmeldingen in twee dagen inderdaad groter is.
(2pt)

2f [6pt] Bereken de kans dat er gedurende een willekeurige week meer dan veertig aanmeldingen binnenkomen.

<u>**Hint**</u>: gebruik de normale benadering van de Poissonverdeling en let op de continuiteitscorrectie!

Stel $\underline{k}$ is het aantal aanmeldingen gedurende een willekeurige week. We willen de kans $P(\underline{k} > 40)$ berekenen. (1pt)

Merk op $\mu = \lambda t = 4{,}455 * 7 \approx 31{,}1844$ (1pt)

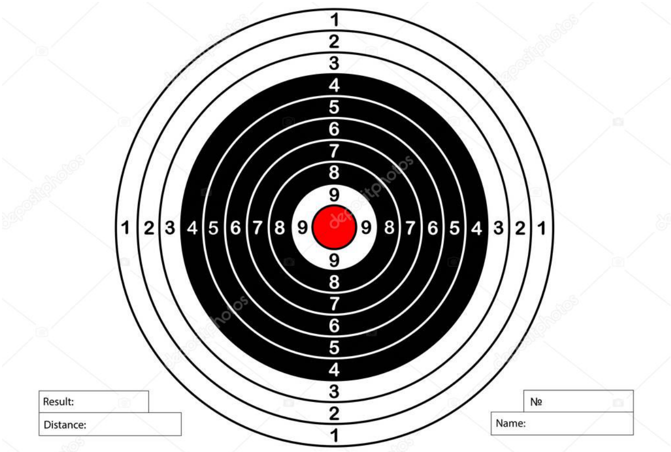
Omdat $\mu \geq 10$, kunnen we de Poissonverdeling benaderen met een normale verdeling met verwachtingswaarde $\mu = 31{,}1844$ en standaarddeviatie $\sqrt{\mu} = 5{,}5843$. (1pt)

Laat $\underline{x}$ een kansvariabele zijn die een normale benadering is van $\underline{k}$.
Rekeninghoudend met de continuiteitscorrectie berekenen we $P(\underline{x} > 40.5) = normalcdf(40.5;\ 10^{99};\ 31{,}1844; 5{,}5843) \approx 0.0476$ (3pt)

<u>**Geen continuiteitscorrectie: -2pt**</u>

Opgave 3 (Totaal 25 punten)

Als onderdeel van een schietoefening worden cadetten getest op hun nauwkeurigheid op een doelwit met ringen (score: gehele getal van 0 (doelwit gemist) tot 10 (rode cirkel)).



Van tien cadetten worden de scores van tien schoten bij elkaar opgeteld:

64, 71, 80, 54, 90, 68, 76, 82, 60, 85

3a [8pt] Bereken het gemiddelde en de standaardafwijking van deze steekproef.

Het gemiddelde van deze steekproef is gelijk aan $\overline{x} = \frac{64+71+80+54+90+68+76+82+60+85}{10} = 73$ (2pt)

De variantie van deze steekproef is $s^2 = \frac{1}{n-1}\sum_i (x_i - \overline{x})^2 \approx 134.67$ (2pt)

De standaarddeviatie is hiervan de wortel: $s = \sqrt{134.67} \approx 11.6$ (1+1pt)

Neem aan dat de score van een willekeurig cadet benaderd kan worden door een normale verdeling met gemiddelde en standaarddeviatie zoals berekend bij 3a. Als je je antwoord niet vertrouwt, werk dan door met $\mu = 74$ en $\sigma = 10$.

3b [5pt] Welke score heeft een cadet nodig om bij de top 10% van beste schutters te behoren?

Laat $\underline{k}$ de score zijn van een cadet, dan kan die met een (continue variabele) $\underline{x}$ worden benaderd die normaal verdeeld is met $\mu = 73$ en $\sigma = 11.6$.

De vraag is nu om een linkergrens g te bepalen zodanig dat $P(\underline{x} \geq g) = 0.1$, oftewel $P(\underline{x} \leq g) = 0.9$ (2pt)

Oftewel: $g = InvNorm(0.9; 73; 11.6) \approx 87.866.$ (2pt)

Om bij de top 10% van beste schutters te behoren moet een cadet een score van minimaal 88 behalen (afronden naar boven) (1pt)

3c [4pt] Om een expertkwalificatie te behalen, dient een cadet een score van minimaal 90 te halen. Bereken de kans dat een willekeurige cadet een expertkwalificatie behaalt. Houd hierbij rekening met de continuïteitscorrectie.

De vraag is om de kans $P(\underline{x} \geq 90)$ te berekenen. (1pt)

Rekening houdend met de continuiteitscorrectie betekent dat we eigenlijk de kans $P(\underline{x} \geq 89.5)$ willen bepalen. (1pt)

Deze kans bereken je als $P(\underline{x} \geq 89.5) = normalcdf(89.5, 10^{10}; 73; 11.6) = 0.0657$ (2pt)

3d [8pt] Stel dat we nu de scores van vijf nieuwe cadetten bekijken. Wat is de kans dat minstens één van hen de expertkwalificatie behaalt?

<u>Hints</u>:

- Bekijk de scores $\underline{k}_1, \underline{k}_2, \underline{k}_3, \underline{k}_4, \underline{k}_5$ van de vijf cadetten als vijf onafhankelijke trekkingen uit dezelfde normale verdeling.
- Er geldt dat voor discrete kansvariabelen $\underline{k}_1, \ldots, \underline{k}_n$ dat $P(maximum\ van\ \underline{k}_1, .. \underline{k}_n \geq k) = 1 - P(\underline{k}_1 \leq k-1, \ldots, \underline{k}_n \leq k-1)$

Laat $\underline{k}_1, \underline{k}_2, \underline{k}_3, \underline{k}_4, \underline{k}_5$ de scores zijn van de vijf cadetten, de vraag is om de kans te bepalen op $P(maximum\ van\ \underline{k}_1, \underline{k}_2, \underline{k}_3, \underline{k}_4, \underline{k}_5 \geq 90)$ te berekenen. (1pt)

Deze kans is moeilijk om uit te rekenen, vandaar dat we omdenken:

$$P(maximum\ van\ \underline{k}_1, \underline{k}_2, \underline{k}_3, \underline{k}_4, \underline{k}_5 \geq 90) = 1 - P(maximum\ van\ \underline{k}_1, \underline{k}_2, \underline{k}_3, \underline{k}_4, \underline{k}_5 \leq 89)$$

(1pt)

Vanwege de onafhankelijkheid van de vijf scores geldt:

$$P(maximum\ van\ \underline{k}_1, \underline{k}_2, \underline{k}_3, \underline{k}_4, \underline{k}_5 \leq 89) = P(\underline{k}_1 \leq 89)P(\underline{k}_2 \leq 89) * \ldots * P(\underline{k}_5 \leq 89)$$

(2pt)

Uit 3c) kunnen we halen dat $P(\underline{k} \geq 90) = 0.0657$, dus $P(\underline{k} \leq 89) = 1 - 0.0657 = 0.9343$ (1pt)

Oftewel de kans is gelijk aan $P(maximum\ van\ \underline{k}_1, \underline{k}_2, \underline{k}_3, \underline{k}_4, \underline{k}_5 \leq 89) = 0.9343^5 = 0.7119$ (2pt)

Dus: $P(maximum\ van\ \underline{k}_1, \underline{k}_2, \underline{k}_3, \underline{k}_4, \underline{k}_5 \geq 90) = 1 - P(maximum\ van\ \underline{k}_1, \underline{k}_2, \underline{k}_3, \underline{k}_4, \underline{k}_5 \leq 89) = 1 - 0.7119 = 0.2881$ (1pt)

Opgave 4 (Totaal 20 punten)

Vanaf maandag 16 september 2024 vinden er grenscontroles plaats in Duitsland als reactie op de toegenomen terrorismedreiging en illegale migratie. Na enkele weken is gebleken dat bij 7% van de staande gehouden auto's minstens een van de inzittenden een migrant is zonder visum.

4a [8pt] Bereken de verwachtingswaarde en de standaarddeviatie van het aantal auto's met minstens een illegale migrant in het geval van een steekproef van 100 auto's.

We hebben hier te maken met een aantal experimenten (autocontroles) met een succeskans $p = 0.07$ op het aantreffen van een illegale migrant. We hebben dus te maken met de binomiale verdeling. (2pt)

De verwachtingswaarde van een binomiaal verdeelde kansvariabele $\underline{x}$ met parameters $n = 100$ en $p = 0.07$ is gelijk aan $E(\underline{x}) = np = 100 * 0.07 = 7$ (2pt)

De variantie van $\underline{x}$ is gelijk aan $Var(\underline{x}) = np(1-p) = 100 * 0.07 * (1 - 0.07) = 6.51$ (2pt)

De standaardafwijking is de wortel hiervan, dus $\sigma_{\underline{x}} = \sqrt{Var(\underline{x})} = \sqrt{6.51} \approx 2.5515$ (2pt)

4b [4pt] Bereken de kans dat pas bij de zesde auto voor het eerst een illegale migrant zich onder de inzittenden bevindt.

Dit betekent dat de eerste vijf controles niet een succes zijn, dit heeft een kans van $(1-p)^5 = 0.93^5 \approx 0.6957$. (2pt)

Daarnaast is de zesde controle wel een succes, dit heeft een kans van $p = 0.07$. (1pt)

Combineren geeft ons $(1-p)^5 p \approx 0.0487$ (1pt)

4c [4pt] Bereken een 95%-voorspellingsinterval van het aantal auto's met minstens een illegale migrant in het geval van een steekproef van 100 auto's.

Hint: gebruik de vuistregel voor 95%-voorspellingsintervallen.

Volgens de vuistregel in het boek kun je een 95%-voorspellingsinterval geven met $[\mu - 2\sigma, \mu + 2\sigma] = [7 - 2 * 2.5515, 7 + 2 * 2.5515] = [1.897, 12.103]$ (4pt)

4d [4pt] Is het 95%-voorspellingsinterval voor het steekproefpercentage smaller of breder als de steekproef wordt uitgebreid naar 200 auto's. Leg uit met een formule.

De formule van de standaarddeviatie voor fracties in de binomiale verdeling is

$\sigma = \sqrt{\frac{p(1-p)}{n}}$ (1pt)

Indien de steekproefgrootte groter wordt, wordt de standaarddeviatie kleiner (2pt)

Hierdoor wordt het interval $[\mu - 2 * \sigma, \mu + 2 * \sigma]$ dus kleiner (1pt)

Bonus (2pt): Met welke factor wordt het interval groter of kleiner?

Bonus: eerst is het interval $4 * \sigma$ breed. Door verdubbeling van de steekproefgrootte verandert de standaardafwijking met een factor $\frac{1}{\sqrt{2}}$, en dus ook het interval zelf (1+1pt)