TILASTOMATEMATIIKKA

Harjoitusviikon 2 tehtävien ratkaisut, kevät 2024

Harjoituksen teemoja ovat:

(i) Kertymäfunktio, pistetodennäköisyysfunktio ja tiheysfunktio

(ii) Diskreetti satunnaismuuttuja

(iii) Jatkuva satunnaismuuttuja

(iv) Tyypillisimmät jakaumat, jotka kantavat omaa nimeä

Tehtävien laskemista voi rytmittää esimerkiksi niin, että alkuviikosta käy läpi tehtäviä 1-5, joista tehtävät 1 ja 2 liittyvät satunnaismuuttujien luokitteluun ja loput tehtävät liittyvät diskreetteihin satunnaismuuttujiin. Tehtävät 6-9 liittyvät jatkuviin satunnaismuuttujiin. Toki myös muunlainen rytmittäminen on mahdollista. Alkuviikon harjoituksissa suositellaan laskettavaksi **tehtävät 2 ja 3** ja loppuviikon harjoituksissa **tehtävät 6 ja 8**.

1. Ilmoita kussakin seuraavista tapauksista, onko kyse jatkuvasta vai diskreetistä satunnaismuuttujasta. Määrää myös satunnaismuuttujan arvojoukko, mikäli se on mahdollista. Voidaanko mallintamisessa käyttää hyväksi jotain tunnettua jakaumaa?

   a) Vikojen lukumäärä neliömetrillä satunnaisesti valitussa paperirullassa.

   b) Kemikaalin konsentraatio liuoksessa.

   c) Liian pitkien pulttien osuus satunnaisesti valitussa pultteja sisältävässä laatikossa.

   d) Virheiden lukumäärä 1000 satunnaisesti valitussa rivissä ohjelmointikoodia.

   e) Satunnaisesti valitun metallilevyn murtolujuus.

   f) Elektronisen komponentin elinikä.

Ratkaisu:

a) Vikojen lukumäärä on ei-negatiivinen kokonaisluku, sillä vikoja voi olla nolla, yksi, kaksi,... kappaletta, joten kyseessä on *diskreetti sm.*
Periaatteessa vikojen lukumäärälle X ei ole ylärajaa, joten arvojoukokoksi voidaan ottaa $S_X = \mathbb{N}_0 = \{0,1,2,\dots\}$.
Vikojen lukumäärä jotakin yksikköä (tässä tapauksessa neliömetriä) kohden on tyypillinen Poisson-jakauman sovelluskohde, joten muuttujan X voidaan olettaa noudattavan Poisson-jakaumaa. On syytä korostaa, että Poisson-jakauman käyttäminen tarkoittaa matemaattisen mallin muodostamista ongelmalle. Todellinen vikojen lukumäärä voi periaatteessa olla mitä hyvänsä. Jos kuitenkin Poisson-jakauman käyttö antaa havaintojen kanssa sopusoinnussa olevia tuloksia, on sen käyttö perusteltua.

b) Vastaus riippuu konsentraation X määritelmästä. Jos konsentraatiolla tarkoitetaan liuenneen aineen massan suhdetta liuoksen tilavuuteen, voi konsentraatio saada periaatteessa minkä tahansa ei-negatiivisen reaalilukuarvon, jolloin arvojoukoksi voidaan valita $S_X = [0,\infty[$. Tokihan käytännössä konsentraatiolle täytyy olla jokin yläraja, mutta koska yläarajaa ei tiedetä, voidaan arvojoukoksi ottaa rajoittamaton väli $[0,\infty[$. Tällöin kyseessä on *jatkuva satunnaismuuttuja.*
Konsentraation voidaan olettaa noudattavan normaalijakaumaa, vaikka normaalijakauma saa kaikki reaalilukuarvot, kun taas konsentraatio ei voi olla negatiivinen. Jälleen normaalijakauman valinnalla muodostetaan matemaattinen malli, joka enemmän tai vähemmän kuvaa todellista tilannetta.

c) Koska laatikon kokoa N ei tunneta, voi pulttien osuus saada periaatteessa minkä tahansa ei-negatiivisen rationaalilukuarvon, joten pulttien osuutta kuvaavan sm:n X arvojoukoksi voidaan valita

$$S_X = \left\{ \frac{m}{n} \,:\, 0 \leq m \in \mathbb{Z}, 0 < n \in \mathbb{Z} \right\}.$$

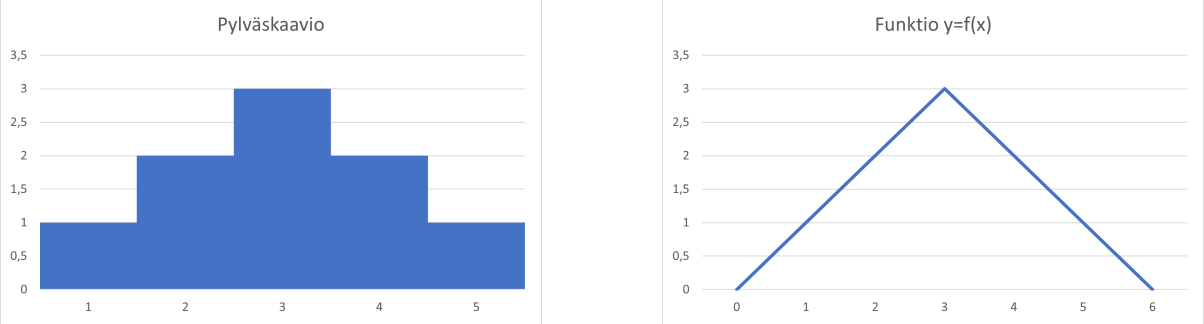
Tällöin kyseessä on *diskreetti sm.*
Jakaumasta emme voi periaatteessa sanoa mitään. Mutta jos laatikon koko tunnetaan, niin viallisten pulttien lukumäärä laatikossa noudattaa binomijakaumaa.

d) Koska virheiden lukumäärä voi olla periaatteessa mikä tahansa ei-negatiivinen kokonaisluku, voidaan tässäkin käyttää Poisson-jakaumaa, jolloin $S_X = \mathbb{N}_0$ ja kyseessä on *diskreetti sm.*
Tarkkaan ottaen Poisson-jakauman käyttö ei ole oikein, sillä yhdellä rivillä ei voi olla loputtomasti merkkejä. Toisakseen, koska eri riveillä voi olla eri määrä merkkejä, emme voi käyttää binomijakaumaa. Toisaalta, koska rivejä on kohtuullisen paljon, kannattaisi käytännössä käyttää Poisson-jakaumaa, vaikka kullakin rivillä olisi sama määrä merkkejä ja siten binomijakauma olisi oikea jakaumamalli.

e) Satunnaisesti valitun metallilevyn murtolujuus voi olla periaatteessa mikä tahansa ei-negatiivinen reaaliluku, jolloin kyseessä olisi *jatkuva satunnaismuuttuja.* Kuten b)-kohdassa, jakaumamallina voisimme käyttää normaalijakaumaa, vaikka se ei sovikaan yhteen (teoreettisen) arvojoukon $[0,\infty[$ kanssa.

f) Elinikä X voi olla periaatteessa mikä tahansa ei-negatiivinen reaaliluku, joten arvojoukoksi voidaan valita $S_X = [0,\infty[$ ja siten kyseessä on *jatkuva sm.* Yksinkertaisin eliniän jakaumamalli on eksponenttijakauma.

2. Kuvaan 1 on piirretty satunnaismuuttujiin liittyviä kuvioita.



Kuva 1: Satunnaismuuttujiin liittyviä kuvioita

a) Voivatko kuviot esittää jotain satunnaismuuttujaa sellaisenaan? Jos ei, niin miten voit skaalata kuvioita, että niistä tulee keskeisiä satunnaismuuttujia kuvaavia funktioita? Mikä on näin saatujen satunnaismuuttujien arvojoukko ja tyyppi?

b) Millä todennäköisyydellä satunnaismuuttujat ovat suurempaa kuin 2?

c) Mikä on kummankin satunnaismuuttujan todennäköisin arvo?

Ratkaisu:

a) Pylväskaavio voisi viitata esimerkiksi lukumääriin. Todennäköisyysjakauma siitä saadaan, kun lasketaan lukumäärien suhteelliset osuudet. Välttämättä ei tarvita lukumääriä, vaan todetaan, että annettu pylväskaavio on todennäköisyyshistogrammiin verrannollinen. Tarkastellaan siis *diskreettiä satunnaismuuttujaa* X, jonka arvojoukko on $S_X = \{1,2,3,4,5\}$. Koska pylväiden yhteenlaskettu pinta-ala on $1+2+3+2+1 = 9$, saadaan kaaviosta todennäköisyyshistogrammi, kun korkeudet jaetaan luvulla 9. Saadaan diskreetti jakauma

x	1	2	3	4	5
$\mathbb{P}(X=x)$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{2}{9}$	$\frac{1}{9}$

Vastaavasti oikeanpuoleinen funktio voisi esittää jonkin satunnaismuuttujan tiheysfunktiota. Koska kolmion pinta-ala on $\frac{1}{2}\cdot 6\cdot 3 = 9$, saadaan funktiosta tiheysfunktio, kun funktiota skaalataan luvulla $1/9$. Koska annetun funktion lauseke on

$$f(x) = \begin{cases} x, & 0 \le x \le 3, \\ 6-x, & 3 \le x \le 6, \\ 0, & \text{muulloin}, \end{cases}$$

saadaan tästä *jatkuvan satunnaismuuttujan* X tiheysfunktio

$$f_X(x) = \begin{cases} \frac{1}{9}x, & 0 \le x \le 3, \\ \frac{1}{9}(6-x), & 3 \le x \le 6, \\ 0, & \text{muulloin}. \end{cases}$$

Arvojoukko S_X on reaalilukuväli $[0,6]$.

b) Kysytään todennäköisyyttä $\mathbb{P}(X > 2)$. Diskreetin muuttujan tapauksessa tämä on sama kuin $\mathbb{P}(X \ge 3) = 1 - \mathbb{P}(X < 3) = 1 - \mathbb{P}(X \le 2)$. Huomaa, että tämä voidaan ratkaista alkuperäisestä kaaviosta suhteen avulla

$$\mathbb{P}(X > 2) = 1 - \mathbb{P}(X \le 2) = 1 - \frac{1+2}{1+2+3+2+1} = 1 - \frac{3}{9} = \frac{2}{3}.$$

Sama tulos toki saadaan myös muodostamastamme jakaumasta, sillä

$$\mathbb{P}(X > 2) = 1 - \mathbb{P}(X \leq 2) = 1 - (\mathbb{P}(X = 1) + \mathbb{P}(X = 2)) = 1 - (\tfrac{1}{9} + \tfrac{2}{9}) = \tfrac{2}{3}.$$

Vastaavasti jatkuvalle muuttujalle saadaan alkuperäisestä funktiosta suhteen avulla

$$\mathbb{P}(X > 2) = 1 - \mathbb{P}(X \leq 2) = 1 - \frac{\frac{1}{2} \cdot 2 \cdot 2}{\frac{1}{2} \cdot 6 \cdot 3} = 1 - \tfrac{2}{9} = \tfrac{7}{9},$$

missä suhde on saatu vertaamalla välille [0,2] muodostuvan kolmion alaa funktion määräämän ison kolmion alaan. Samaan lopputulokseen päästään toki myös integroimalla

$$\mathbb{P}(X > 2) = 1 - \mathbb{P}(X \leq 2) = 1 - \int_0^2 \tfrac{1}{9} x \mathrm{d}x = 1 - \tfrac{1}{2 \cdot 9} x^2 \Big|_{x=0}^{2} = 1 - \tfrac{2}{9} = \tfrac{7}{9}.$$

c) Diskreetille jakaumalle todennäköisin arvo lienee se, jonka pistetodennäköisyys on kaikkein suurin. Kuvan perusteella se on $X = 3$. Jatkuvalle jakaumalle taasen samanlainen luokittelu ei ole mahdollista, sillä jokaisen yksittäisen pisteen todennäköisyys on nolla. Sen sijaan lienee ihan järkevää tarkastella tiheysfunktion maksimikohtaa. Tämäkin voidaan määrätä alkuperäisestä funktiosta, joten ”todennäköisin” arvo on $X = 3$.
Edellä määrätyille arvoille käytetään nimitystä *moodi*.

3. Eräällä tehtaalla valmistetuista mikrosiruista 2 % on viallisia. Poimitaan satunnaisotannalla 100 mikrosirua tarkastukseen. Millä todennäköisyydellä tarkastuksessa löydetään 5 viallista mikrosirua?

   a) Laske todennäköisyys tilanteeseen sopivan satunnaismuuttujan avulla.

   b) Kuinka monta viallista mikrosirua löydetään keskimäärin? Millä todennäköisyydellä viallisten määrä on korkeintaan 4?

   c) Laske kysytty todennäköisyys Poisson-jakauman avulla.

   d) Simuloi kysyttyä todennäköisyyttä Excelillä. Käytä BINOM.INV- ja COUNTIF-funktioita sekä kurssilla esitettyä käänteismuunnosmenetelmää. Simuloi esimerkiksi 10000 näytettä binomijakaumasta. Voit ottaa mallia vaikkapa videosta

   https://www.youtube.com/watch?v=5uF0L6nImn4

   kohdasta 23:15 eteenpäin.

Ratkaisu:
Olkoon

$$X = \text{"viallisten mikrosirujen lukumäärä tarkastuksessa"}.$$

Kysytään todennäköisyyttä $\mathbb{P}(X = 5)$.

a) Oletetaan, että mikrosirujen viallisuus on toisistaan riippumatonta. Tällöin X noudattaa binomijakaumaa parametreilla $n = 100$ ja $p = 0.02$, mitä merkitään $X \sim \text{Bin}(100,0.02)$. Kaavakokoelmasta löytyy binomijakauman pistetodennäköisyyksien laskentakaava

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0,1,\ldots,n, \tag{1}$$

jonka mukaan

$$\mathbb{P}(X = 5) = \binom{100}{5} 0.02^5 0.98^{100-2} \approx 0.035,$$

eli tarkastuksessa löydetään 5 viallista mikrosirua noin 3,5 % todennäköisyydellä.

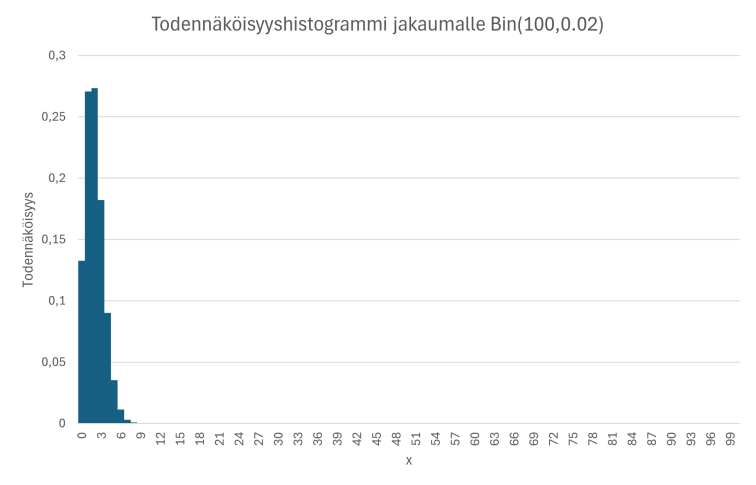
b) Tässä "keskimäärin" on hieman epämääräinen. Tyypillisesti sillä tarkoitetaan odotusarvoa $\mu_X = np = 100 \cdot 0.02 = 2$, mutta siitä enemmän ensi viikolla. Jos jakaumalle $X \sim \text{Bin}(100,0.02)$ piirretään vaikkapa Excelillä todennäköisyyshistogrammi tai katsotaan Excelin laskemia pistetodennäköisyyksiä, niin nähdään, että todennäköisin arvo on $X = 2$, joka vastannee mielikuvaa "keskimääräisestä arvosta". Tämä näyttäisi ainakin tässä tapauksessa yhtyvän odotusarvoon. Kuvaajasta tai laskelmista nähdään, että arvot $X \geq 8$ ovat "hyvin epätodennäköisiä. Lähes kaikki todennäköisyysmassa on keskittynyt välille [0,7]. Tämän perusteella taasen voisi sanoa, että X saa "keskimäärin" arvoja ko. väliltä.
Kysytään todennäköisyyttä

$$\mathbb{P}(X \leq 4) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \mathbb{P}(X = 2) + \mathbb{P}(X = 3) + \mathbb{P}(X = 4),$$

jossa hyödynnettiin sitä, että muuttujan X arvojoukko on $S_X = \{0,1,2,\ldots,100\}$. Vaikka yhteenlaskettavia ei ole kuin 5, jotka kukin voitaisiin laskea pistetodennäköisyyden laskukaavalla (1), hyödynnetään tässä kertymäfunktiota $F_X(x) = \mathbb{P}(X \leq x)$, jonka mukaan kysytty todennäköisyys on

$$\mathbb{P}(0 \leq X \leq 4) = F_X(4) \approx 0.949$$

eli noin 95 %. Kertymäfunktion arvon laskemiseen käytettiin Excelin BINOM.DIST-funktiota. Palataksemme ensimmäiseen kysymykseen, loppuosan perusteella voitaisiin sanoa, että viallisten lukumäärä on "keskimäärin" välillä [0,4], joka on viallisten lukumäärän 95 % todennäköisyysväli.



Kuva 2: Tehtävän 2 b) histogrammi

c) Merkitään $a = np = 2$ ja oletetaan, että $X \sim \text{Poi}(2)$. Tällöin kysytyksi todennäköisyydeksi saadaan kaavakokoelman mukaan

$$\mathbb{P}(X = 5) = \mathrm{e}^{-2}\frac{2^5}{5!} \approx 0.036.$$

Vaihtoehtoisesti todennäköisyys voidaan laskea Excelin POISSON.DIST-funktiolla.

d) Kun meneteltiin ohjeen mukaisesti, löyti eräs kokoa $n = 10000$ oleva simulaatio 366 havaintoa, joissa saatiin arvo 5. Siten todennäköisyydeksi saatiin 0.037 promillen tarkkuudella. Kootaan edellä laskettu taulukoksi

Laskutapa	Binomi	Poisson	simulaatio
$\mathbb{P}(X = 5)$	0.035	0.036	0.037

Eri laskutavat näyttivät antavan likimain samat tulokset.

4. Erittäin laaja havupuualue oli neulaskadon takia jäämässä tuotantotavoitteestaan. Tuottoennusteen tekemiseksi tietty osa metsää jaettiin hehtaarin suuruisiin ruutuihin ja laskettiin neulaskadon vaivaamien puiden lukumäärä ruutua kohti. Laskennan tuloksena havaittiin neulaskadon vaivaamien puiden lukumäärä/ruutu olevan Poisson-jakautuneen ja sellaisten ruutujen, joissa neulaskadon vaivaamia puita ei ollut lainkaan, osuuden olevan 7 %.

   a) Mikä oli oli keskimääräinen neulaskadon vaivaamien puiden lukumäärä ruutua kohti?

   b) Millä todennäköisyydellä ruudussa esiintyi ainakin 2 neulaskadon vaivaamaa puuta?

Ratkaisu: Olkoon $X =$ "neulaskadon vaivaamien puiden lukumäärä/ha". Tehtävänannon mukaan $X \sim \text{Poi}(a)$, missä parametria a ei vielä tiedetä.

a) Poisson-jakauman parametri a ilmoittaa satunnaismuuttujan keskimääräisen arvon. Tehtävänannon mukaan

$$\mathbb{P}(X = 0) = \mathrm{e}^{-a} = 0.07 \Rightarrow a = -\ln 0.07 \approx 2.659,$$

neulaskadon vaivaamia puita on keskimäärin 2.7 kappaletta hehtaaria kohti.

b) Kysytään todennäköisyyttä $\mathbb{P}(X \geq 2)$, joka ainakin funktiolaskimella on helpoin laskea komplementin kautta

$$\mathbb{P}(X \geq 2) = 1 - \mathbb{P}(X < 2) = 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) = 1 - \mathrm{e}^{\ln 0.07}(1 - \ln 0.07) \approx 74\%.$$

5. Valmistajan ilmoituksen mukaan halvoista elektronisista komponenteista AA5 keskimäärin 5 % ei täytä spesifikaatioita. Ostaja osti AA5:ttä monen tuhannen kappaleen erissä ja teki seuraavan päätöksentekomallin erän hyväksymiseksi vastaanottotarkastuksen yhteydessä: Erästä valittiin satunnaisesti 10 komponenttia, jotka testattiin. Jos kaikki testatut komponentit toteuttivat spesifikaatiot, niin erä hyväksyttiin. Jos ainakin 2 testatuista komponenteista ei toteuttanut spesifikaatioita, erä hylättiin. Jos testatuista täsmälleen yksi ei toteuttanut spesifikaatioita, otettiin uusi 10 yksilön satunnainen näyte, ja erä hyväksyttiin, jos uudessa näytteessä kaikki komponentit täyttävät spesifikaatiot. Jos valmistajan väite pitää paikkansa, niin millä todennäköisyydellä

   a) erä hyväksytään ensimmäisen näytteen perusteella?

   b) lopullinen päätös tehdään eli erä hylätään tai hyväksytään ensimmäisen näytteen perusteella?

   c) erä hyväksytään?

Ratkaisu: Olkoon

$$X_i = \text{”spesifikaation täyttävien komponenttien lukumäärä näytteessä } i\text{”}, \quad i = 1{,}2.$$

Koska komponentit toimitetaan monen tuhannen kappaleen erissä, voidaan olettaa, että $X_i \sim \text{Bin}(10{,}0.95)$, $i = 1{,}2$, ja että X_1 ja X_2 ovat riippumattomia.

a) Erä hyväksytään ensimmäisen näytteen perusteella, kun $X_1 = 10$, joten kysytty todennäköisyys on

$$\mathbb{P}(X_1 = 10) = \binom{10}{10} 0.95^{10} \cdot 0.05^{10-10} \approx 60\%.$$

b) Lopullinen päätös tehdään ensimmäisen näytteen perusteella täsmälleen silloin, kun $X_1 \neq 9$, joten kysytty todennäköisyys on

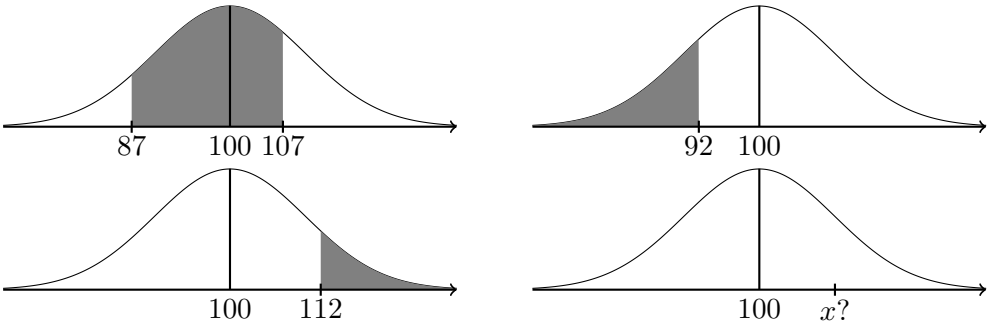
$$\mathbb{P}(X_1 \neq 9) = 1 - \mathbb{P}(X = 9) = 1 - \binom{10}{9} 0.95^{9} \cdot 0.05^{10-9} \approx 68\%.$$

c) Erä hyväksytään joko ensimmäisen tai toisen näytteen perusteella. Kysytään tapahtuman $\{X_1 = 10\} \cup (\{X_1 = 9\} \cap \{X_2 = 10\})$ todennäköisyyttä. Koska yhdisteen tapahtumat ovat erillisiä sekä X_1 ja X_2 ovat riippumattomia, on kysytty todennäköisyys

$$p = \binom{10}{10} 0.95^{10} \cdot 0.05^{0} + \binom{10}{9} 0.95^{9} \cdot 0.05^{1} \cdot \binom{10}{10} 0.95^{10} \cdot 0.05^{0} \approx 79\%.$$

6. Tarkastellaan satunnaismuuttujaa $X \sim N(\mu,10^2)$, jonka tiheysfunktion profiileja on piirretty Kuvaan 3.

   a) Laske väritettyjen alueiden pinta-alat.



Kuva 3: Tehtävän 6 tiheysfunktiot

   b) Missä pisteessä x pätee
   (i)$\mathbb{P}(X \leq x) = 0.93$?
   (ii)$\mathbb{P}(X \leq x) = 0.38$?

   c) Olkoot $X_1,X_2,\ldots,X_{10}$ riippumattomia ja samalla tavalla jakautuneita kuin X. Millä todennäköisyydellä muuttujien maksimi on suurempaa kuin 107? Toisin sanoen, laske todennäköisyys $\mathbb{P}(\max_{1\leq i\leq 10} X_i > 107)$.

Ratkaisu:
Koska tiheysfunktio saavuttaa maksiminsa odotusarvopisteessä, niin kuvaajien perusteella $\mu = 100$.

a) Vasemmassa yläkulmassa oleva väritetty alue vastaa todennäköisyyden $\mathbb{P}(87 \leq X \leq 107)$ laskemista. Tarkastellaan standardoitua muuttujaa $Z = \frac{X-\mathbb{E}(X)}{\sigma_X} = \frac{X-100}{10} \sim N(0,1)$. Todennäköisyydeksi saadaan

$$\begin{aligned}
\mathbb{P}(87 \leq X \leq 107) &= \mathbb{P}\left(\frac{87-100}{10} \leq \frac{X-\mathbb{E}(X)}{\sigma_X} \leq \frac{107-100}{10}\right) && \text{(standardointi)}\\
&= \mathbb{P}(-1.3 < Z \leq 0.7) && (X \text{ on jatkuva sm.})\\
&= \Phi(0.7) - \Phi(-1.3) && (Z \sim N(0,1))\\
&= \Phi(0.7) - (1 - \Phi(1.3)) && \text{(symmetria-ominaisuus)}\\
&\approx 0.7580 - (1 - 0.9032) && \text{(taulukosta)}\\
&\approx 66\,\%.
\end{aligned}$$

Oikealla ylhäällä olevassa kuvassa kysytään todennäköisyyttä $\mathbb{P}(X \leq 92)$. Kuten yllä saadaan

$$\mathbb{P}(X \leq 92) = \mathbb{P}\left(\frac{X-100}{10} \leq \frac{92-100}{10}\right) = \Phi(-0.8) = 1 - \Phi(0.8) \approx= 1 - 0.7881 \approx 21\,\%.$$

Edelleen vasemmalla alhaalla olevalle kuvalle saadaan todennäköisyydeksi

$$\mathbb{P}(X > 112) = 1 - \mathbb{P}(X \leq 112) = 1 - \mathbb{P}\left(Z \leq \frac{112-100}{10}\right) = 1 - \Phi(1.2) \approx= 1 - 0.8849 \approx 12\,\%.$$

b) Jälleen standardoimalla saadaan

$$\mathbb{P}(X \leq x) = \mathbb{P}\left(\frac{X-100}{10} \leq \frac{x-100}{10}\right) = 0.93$$

Taulukosta tai numeerisesti Excelin NORM.S.INV-funktiolla saadaan $z = \frac{x-100}{10} = 1.48$, josta $x = 114.8$.
Jälkimmäiselle pisteelle taas taulukoita käyttämällä täytyy hyödyntää symmetria-ominaisuutta, jonka mukaan

$$\mathbb{P}(X \leq x) = \mathbb{P}\left(Z \leq \frac{x-100}{10}\right) = 0.38 \Leftrightarrow \mathbb{P}\left(Z \geq \frac{x-100}{10}\right) = \Phi\left(-\frac{x-100}{10}\right) = 0.62.$$

Ekvivalenssissa on hyödynnetty komplementtia ja viimeistä edellisessä yhtäsuuruudessa on käytetty symmetria-ominaisuutta, jonka mukaan pisteen $z = \frac{x-100}{10}$ oikealle puolelle jää sama määrä todennäköisyysmassaa kuin pisteen $-z$ vasemmalle puolelle, minkä taasen antaa kertymäfunktio.
Taulukosta tai numeerisesti Excelin NORM.S.INV-funktiolla saadaan $-z = 0.31$, josta $x = 100 - 10 \cdot 0.31 = 96.9$. Itse asiassa NORM.S.INV-funktio antaa z-pisteen suoraan ilman symmetria-ominaisuutta. Komento "=NORM.INV(0,38;100;10)" antaa kyllä kysytyn pisteen suoraan, mutta kannattaa opetella standardoinnin ja symmetria-ominaisuuden käyttö tässäkin yhteydessä.

c) Komplementin kautta saadaan

$$\mathbb{P}(\max_{1 \leq i \leq 10} X_i > 107) = 1 - \mathbb{P}(\max_{1 \leq i \leq 10} X_i \leq 107).$$

Koska muuttujien maksimi on pienempää (tai yhtäsuurta) kuin 107 tasan silloin, kun kaikki muuttujat ovat pienempää (tai yhtäsuurta) kuin 107, saadaan samalla tavalla kuin a)-kohdassa

$$\mathbb{P}(\max_{1 \leq i \leq 10} X_i > 107) = 1 - (\mathbb{P}(X_1 \leq 107))^{10} = 1 - (\Phi(0.7))^{10} \approx 1 - 0.7580^{10} \approx 94\,\%,$$

sillä muuttujat ovat riippumattomat ja noudattavat samaa jakaumaa $N(100,10^2)$.

7. Eloonjäämisfunktio S (*survival function*) ilmoittaa todennäköisyyden, että henkilö on elossa tietyn ajan kuluttua. Oletetaan, että syöpäpotilaan eloonjäämisfunktio on muotoa $S(t) = 1 - F_X(t)$, missä t on syöpädiagnoosista kulunut aika vuosina ja X on elinajan ilmoittava satunnaismuuttuja, jonka oletetaan noudattavan Weibull-jakaumaa parametreilla $\alpha = 0.98$ ja $\beta = 0.30$.

   a) Esitä Weibull-jakauman tiheysfunktion lauseke (löytyy esimerkiksi luentomonisteesta).

   b) Laske todennäköisyys, että henkilö on elossa 5 vuotta syöpädiagnoosin jälkeen.

   c) Mikäli henkilö on elossa 5 vuotta syöpädiagnoosin jälkeen, niin millä todennäköisyydellä hän elää vielä 5 vuotta?

Ratkaisu:

a) Weibull-jakauman tiheysfunktion lauseke löytyy esimerkiksi luentomonisteen sivulta 41 tai vaikkapa Wikipedia-artikkelista

https://en.wikipedia.org/wiki/Weibull_distribution

Tiheysfunktio on

$$f(t) = \begin{cases} \alpha\beta t^{\beta-1} e^{-\alpha t^\beta}, & t \geq 0, \\ 0, & t < 0. \end{cases}$$

Huomaa, että Wikipediassa ja luentomonisteessa esitetty tiheysfunktion määritelmä poikkevat toisistaan parametrien osalta. Wikipediassa esitetyn tiheysfunktion parametrit ovat λ ja k, joiden yhteys yllä esitettyihin parametreihin α ja β on

$$\beta = k \quad \text{ja} \quad \alpha = \frac{1}{\lambda^\beta}.$$

Vaikka tiheysfunktion lauseke saattaa näyttää aluksi kummalliselta, huomioi, että eksponenttifunktion edessä on merkkiä vaille eksponentin derivaatta, joten tiheysfunktio on helppo integroida ja kertymäfunktioksi saadaan suoraan

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t = \int_{0}^{x} f(t)\mathrm{d}t = \begin{cases} 1 - e^{-\alpha x^\beta}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Weibull-jakauma on eksponenttijakauman yleistys, joka ottaa huomioon, että komponentin (tässä tapauksessa ihmisen) vioittumistodennäköisyys voi muuttua iän myötä.

b) Kysytään todennäköisyyttä $\mathbb{P}(X > 5)$, joka voidaan laskea kertymäfunktion avulla seuraavasti

$$\mathbb{P}(X > 5) = 1 - \mathbb{P}(X \leq 5) = 1 - F(5) = S(5) = e^{-0.98 \cdot 5^{0.30}} \approx 0.20 = 20\%.$$

c) Nyt kysytään ehdollista todennäköisyyttä

$$\mathbb{P}(X > 10 | X > 5) = \frac{\mathbb{P}(''X > 10 \,\text{ja}\, X > 5'')}{\mathbb{P}(X > 5)} = \frac{\mathbb{P}(X > 10)}{\mathbb{P}(X > 5)} = \frac{S(10)}{S(5)} \approx 0.69 = 69\%.$$

8. Oletetaan, että komponentin elinaika on eksponenttijakautunut ja että elinajan odotusarvo on 20 päivää.

   a) Esitä eksponenttijakauman tiheysfunktion ja kertymäfunktion lauseke (löytyy esimerkiksi luentomonisteesta). Miten parametri liittyy odotusarvoon?

   b) Millä todennäköisyydellä komponentti kestää vähintään 30 päivää?

   c) Jos komponentti on kestänyt 30 päivää, niin millä todennäköisyydellä komponentti kestää vielä toiset 30 päivää?

   d) Varastossa on 10000 komponenttia. Kuinka monta komponenttia keskimäärin kestää vähintään 30 päivää?

Ratkaisu:

a) Satunnaismuuttujan $X \sim \mathrm{Exp}(\lambda)$ tiheysfunktion kaava

$$f_X(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x}, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

löytyy esimerkiksi luentomonisteesta tai vaikkapa Wikipediasta. Kertymäfunktio löytyy kyllä samaisista lähteistä, mutta lasketaan se nyt harjoituksen vuoksi integroimalla

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\mathrm{d}y = \begin{cases} \int_0^x \lambda \mathrm{e}^{-\lambda y}\mathrm{d}y = -\mathrm{e}^{-\lambda y}\Big|_{y=0}^{x} = 1 - \mathrm{e}^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Odotusarvo saadaan parametrin käänteislukuna. Kaava $\mathbb{E}(X) = \frac{1}{\lambda}$ löytyy myös muun muassa kaavakokoelmasta.

b) Tässä tapauksessa a)-kohdan mukainen parametri on $\lambda = \frac{1}{20}$. Kysytään todennäköisyyttä $\mathbb{P}(X \geq 30)$, joka voidaan laskea kahdella tavalla. Ensimmäinen tapa on suora integrointi

$$\mathbb{P}(X \geq 30) = \int_{30}^{\infty} \tfrac{1}{20}\mathrm{e}^{-\frac{1}{20}x}\mathrm{d}x = -\mathrm{e}^{-\frac{1}{20}x}\Big|_{x=30}^{\infty} = \mathrm{e}^{-\frac{3}{2}} \approx 22\,\%.$$

Toinen tapa on kertymäfunktion käyttö, jonka voi ottaa annettuna, jos sen sattuu tuntemaan. Suosittelen kuitenkin käymään integroinnin itse läpi. Nimittäin, koska jatkuvalla jakaumalla yksittäisen pisteen todennäköisyys on nolla, niin ”komplementin kautta” saadaan

$$\begin{aligned} \mathbb{P}(X \geq 30) &= 1 - \mathbb{P}(X \leq 30) = 1 - F_X(30) = 1 - \int_{-\infty}^{30} f_X(x)\mathrm{d}x = 1 - \int_0^{30} \tfrac{1}{20}\mathrm{e}^{-\frac{1}{20}x}\mathrm{d}x \\ &= 1 + \mathrm{e}^{-\frac{1}{20}x}\Big|_{x=0}^{30} = 1 + (\mathrm{e}^{-\frac{3}{2}} - \mathrm{e}^{0}) = \mathrm{e}^{-\frac{3}{2}} \approx 0.22. \end{aligned}$$

c) Kysytään todennäköisyyttä $\mathbb{P}(X \geq 30 + 30 \,|\, X \geq 30)$. Ehdollisen todennäköisyyden määritelmän mukaan

$$\mathbb{P}(X \geq 60 \,|\, X \geq 30) = \frac{\mathbb{P}(\{X \geq 60\} \cap \{X \geq 30\})}{\mathbb{P}(X \geq 30)} = \frac{\mathbb{P}(X \geq 60)}{\mathbb{P}(X \geq 30)} = \frac{\mathrm{e}^{-\frac{6}{2}}}{\mathrm{e}^{-\frac{3}{2}}} = \mathrm{e}^{-\frac{3}{2}} \approx 0.22.$$

Saatiin siis $\mathbb{P}(X \geq 30 + 30 \,|\, X \geq 30) = \mathbb{P}(X \geq 30)$. Tämä pätee yleisemminkin. Nimittäin $\mathbb{P}(X \geq x + h \,|\, X \geq x) = \mathbb{P}(X \geq h)$ kaikilla x,$h \geq 0$. Tätä ominaisuutta kutsutaan *muistinmenetysominaisuudeksi*.

d) Olkoon

$$Y = \text{"niiden komponenttien lukumäärä, jotka kestävät vähintään 30 päivää"}.$$

Jos komponenttien vikaantuminen on toisistaan riippumatonta, niin $Y \sim \text{Bin}(10000,p)$, missä $p = \mathbb{P}(X \geq 30)$. Kysytään odotusarvoa $\mathbb{E}(Y) = 10000p$, jos "keskimäärin" tulkitaan odotusarvona, mutta siitä enemmän ensi viikolla. Edellä laskettiin p, joten odotusarvoksi saadaan noin 2200 sadan komponentin tarkkuudella. Tämän voi ajatella myös ns. arkijärjellä niin, että jos yksittäinen komponentti kestää vähintään 30 päivää todennäköisyydellä 0.22, niin 10000 komponentin joukossa keskimäärin $10000 \cdot 0.22 = 2200$ komponenttia kestää vähintään 30 päivää. Samalla tehtiin itse asiassa heuristinen perustelu binomijakauman odotusarvolle.

9. Valmistetaan metallikuulia, joiden halkaisijaa X pystytään kontrolloimaan. Mikä on kuulien tilavuuden jakauma (kertymäfunktio ja tiheysfunktio), kun oletetaan, että

   a) $X \sim \text{Tas}(a,b)$?

   b) $X \sim N(\mu,\sigma^2)$?

Ratkaisu: Pallon tilavuus V riippuu halkaisijasta X kaavalla

$$V = \frac{1}{6}\pi X^3.$$

Tilavuus V on siis sm:n X muunnos ja on ensin selvitettävä

$$F_V(v) = \mathbb{P}(V \le v).$$

a) Koska ainoastaan positiivinen halkaisija on järkevä, oletetaan, että $0 < a < b$. Tällöin $\mathbb{P}(V > 0) = 1$ ja saadaan

$$F_V(v) = \mathbb{P}(V \le v) = \mathbb{P}\left(X \le \sqrt[3]{\frac{6v}{\pi}}\right)$$

kaikilla $v > 0$. Edelleen

$$a < \sqrt[3]{\frac{6v}{\pi}} < b \Leftrightarrow \frac{1}{6}\pi a^3 < v < \frac{1}{6}\pi b^3,$$

jonka mukaan

$$F_V(v) = \int_a^{\sqrt[3]{\frac{6v}{\pi}}} \frac{1}{b-a}\mathrm{d}x = \frac{1}{b-a}\left(\sqrt[3]{\frac{6v}{\pi}} - a\right), \quad \text{kun } \frac{1}{6}\pi a^3 < v < \frac{1}{6}\pi b^3.$$

Derivoimalla saadaan tiheysfunktioksi

$$f_V(v) = \frac{\mathrm{d}}{\mathrm{d}v}F_V(v) = \frac{1}{b-a}\sqrt[3]{\frac{2}{9\pi}}v^{-2/3}, \quad \text{kun } \frac{1}{6}\pi a^3 < v < \frac{1}{6}\pi b^3.$$

b) Nyt $X \sim N(\mu,\sigma^2)$, joten X voi saada myös negatiivisia arvoja, mikä ei tietenkään ole järkevää sovelluskohde huomioiden. Sen vuoksi oletetaan, että $\mu \gg 0$, jolloin $\mathbb{P}(X < 0) \approx 0$ ja siten $\mathbb{P}(V < 0) \approx 0$. Näin voidaan tehdä esimerkiksi siirtymällä sellaisiin mittayksiköihin, että halkaisija on kaukana nollasta.
Samalla tavalla kuin a)-kohdassa saadaan

$$F_V(v) = \mathbb{P}(V \le v) = \mathbb{P}\left(X \le \sqrt[3]{\frac{6v}{\pi}}\right).$$

Koska X on normaalijakautunut, niin standardisoimalla $Z = \frac{X-\mu}{\sigma}$ saadaan

$$F_V(v) = \mathbb{P}\left(Z \le \frac{\sqrt[3]{\frac{6v}{\pi}} - \mu}{\sigma}\right) = \Phi\left(\frac{\sqrt[3]{\frac{6v}{\pi}} - \mu}{\sigma}\right),$$

missä merkintä Φ tarkoittaa standardisoidun normaalijakauman kertymäfunktiota kuten tavallisesti. Edellisestä derivoimalla saadaan

$$f_V(v) = \frac{\mathrm{d}}{\mathrm{d}v}F_V(v) = \frac{1}{\sigma}\sqrt[3]{\frac{2}{9\pi}}v^{-2/3}\varphi\left(\frac{\sqrt[3]{\frac{6v}{\pi}} - \mu}{\sigma}\right),$$

missä

$$\varphi(z) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{z^2}{2}}$$

on normaalijakauman $Z \sim N(0,1)$ tiheysfunktio.
Yllä olevalla mallilla todennäköisyys negatiiviselle tilavuudelle ei ole nolla, mutta oletettiin, että se on häviävän pieni.