Luku 4

Diskreetit jakaumat

Diskreetti satunnaismuuttuja määriteltiin alaluvussa 2.5. Olemme jo edellisissä luvuissa käsitelleet hypergeometrista jakaumaa (alaluku 2.6.1), binomijakaumaa (alaluvut 2.8 ja 3.6) ja sen erikoistapauksena Bernoullin jakaumaa sekä diskreettiä tasajakaumaa (alaluku 2.5.4), jotka kaikki ovat esimerkkejä *diskreeteistä jakaumista.*

4.1 Diskreetti satunnaismuuttuja

Määritelmä 4.1 Otosavaruudessa Ω määritelty satunnaismuuttuja X on diskreetti, jos sen arvojoukko $S \subset \mathbb{R}$ on numeroituva ja $P(X \in S) = 1$. Joukon S pisteillä on positiivinen todennäköisyys ja ne ovat X:n kertymäfunktion F *hyppypisteitä* ja näiden pisteiden todennäköisyydet ovat F:n hyppyjä.

Määritellään nyt yksinkertainen *hyppyfunktio* $\varepsilon(x)$ seuraavasti:

$$\varepsilon(x) = \begin{cases} 1, & x \geq 0; \\ 0, & x < 0. \end{cases}$$

Olkoon X:n arvoalue $S = \{1, 2, 3, \ldots\}$ ja $P(x = i) = p_i$, $i \geq 1$. Silloin X:n kertymäfunktio $F(X)$ voidaan kirjoittaa muodossa

$$F(x) = \sum_{i=1}^{\infty} p_i \, \varepsilon(x - i). \tag{4.1.1}$$

Vaikka usein tarkastelemme vain kokonaislukuarvoisia satunnaismuuttujia, se ei ole teoreettiselta kannalta oleellinen rajoitus. Olkoon $S^* = \{x_1, x_2, x_3, \ldots\}$ diskreetin satunnaismuuttujan arvojoukko. Silloin joukkojen S ja S^* välillä on bijektiivinen vastaavuus $g(x_i) = i$ ja $P(X = x_i) = P\big(g(X) = i\big)$, joten voimme aina tarvittaessa siirtyä tarkastelemaan vastaavaa kokonaislukuarvoista satunnaismuuttujaa.

Esimerkki 4.1 Yksinkertaisin satunnaismuuttuja X on sellainen, jonka arvoalue $S = \{c\}$ on yksi piste, jolloin $P(X = c) = 1$. Silloin X:n kertymäfunktio on

$$F(x) = \varepsilon(x-c) = \begin{cases} 1, & x \geq c; \\ 0, & x < c. \end{cases}$$

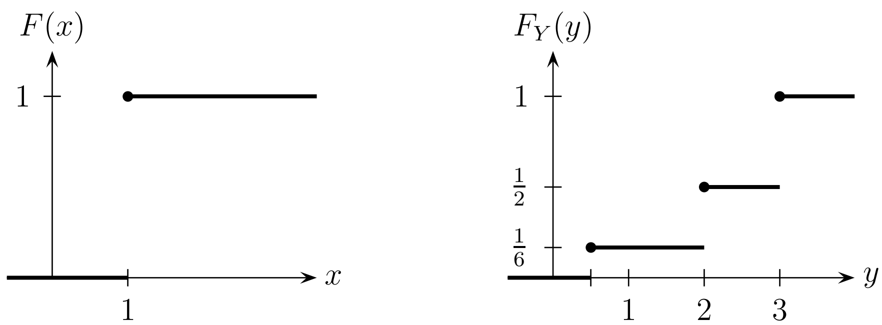
Olkoon Y:n todennäköisyysfunktio

$$P\left(Y = \tfrac{1}{2}\right) = \tfrac{1}{6}, \qquad P(Y = 2) = \tfrac{1}{3} \qquad \text{ja} \qquad P(Y = 3) = \tfrac{1}{2}.$$

Silloin Y:n kertymäfunktio on

$$F_Y(y) = \tfrac{1}{6}\,\varepsilon\left(y - \tfrac{1}{2}\right) + \tfrac{1}{3}\,\varepsilon(y-2) + \tfrac{1}{2}\,\varepsilon(y-3).$$

□



Kuvio 4.1. Funktioiden $F(x) = \varepsilon(x-1)$ ja $F_Y(y)$ kuvaajat.

Esimerkki 4.2 Hatussa on N arpalippua, jotka on numeroitu juoksevasti ykkösestä lähtien. Valitaan hatusta arpa satunnaisesti palauttaen n kertaa ja merkitään valittujen arpojen numerot muistiin. Olkoon X suurin valittujen arpojen numeroista. Silloin $P(X \leq r) = (r/N)^n$ ja

$$\begin{aligned} P(X = r) &= P(X \leq r) - P(X \leq r-1) \\ &= \left(\frac{r}{N}\right)^n - \left(\frac{r-1}{N}\right)^n. \end{aligned}$$

Määritelmän mukaan X:n odotusarvo on

$$\begin{aligned} E(X) &= N^{-n} \sum_{r=1}^{N} [r^n - (r-1)^n] r \\ &= N^{-n} \sum_{r=1}^{N} \left[r^{n+1} - (r-1)^n r\right] \\ &= N^{-n} \sum_{r=1}^{N} \left[r^{n+1} - (r-1)^n\big((r-1)+1\big)\right] \end{aligned}$$

$$
\begin{aligned}
&= N^{-n} \sum_{r=1}^{N} \bigl[r^{n+1} - (r-1)^{n+1} - (r-1)^n \bigr] \\
&= N^{-n} \Bigl[N^{n+1} - \sum_{r=1}^{N} (r-1)^n \Bigr].
\end{aligned}
$$

□

4.2 Bernoullin kokeet ja binomijakauma

Alaluvussa 2.8 binomijakauma esiteltiin tarkastelemalla otantaa palauttaen ja alaluvussa 3.6 binomijakauma liitettiin Bernoullin kokeisiin. Bernoullin koe on satunnaiskoe, jolla on täsmälleen kaksi toisensa poissulkevaa tulosvaihtoehtoa (onnistuminen ja epäonnistuminen — lyhyesti O ja E). Esimerkiksi mielipidetiedustelussa henkilö kannattaa tai ei kannata ehdokasta, laatukontrollissa tuote on virheetön tai viallinen, hoidon tuloksena potilas paranee tai ei parane.

Satunnaismuuttuja X noudattaa *Bernoullin jakaumaa,* kun

$$
X = \begin{cases} 1 & \text{todennäköisyydellä } p, \\ 0 & \text{todennäköisyydellä } 1-p, \end{cases} \tag{4.2.1}
$$

missä $0 \le p \le 1$. Nyt siis X on 'onnistumisen' indikaattorifunktio. Onnistumistodennäköisyys on $P(X = 1) = p$ ja vastaavasti epäonnistumisen todennäköisyys on $P(X = 0) = 1 - p$, jota merkitään usein $q = 1 - p$. Bernoullin jakaumaa noudattavan satunnaismuuttujan X odotusarvo ja varianssi ovat

$$
E(X) = p \qquad \text{ja} \qquad \mathrm{Var}(X) = pq,
$$

sillä

$$
E(X) = p \cdot 1 + q \cdot 0 = p, \qquad E(X^2) = p \cdot 1^2 + q \cdot 0^2 = p
$$

ja

$$
\mathrm{Var}(X) = E(X^2) - [E(X)]^2 = p - p^2 = p(1-p) = pq.
$$

Merkitsemme $X \sim \mathrm{Ber}(p)$, kun X noudattaa Bernoullin jakaumaa, jonka odotusarvo on p.

Jos $X \sim \mathrm{Ber}(p)$, niin X:n kertymäfunktio on

$$
F(x) = (1-p)\,\varepsilon(x) + p\,\varepsilon(x-1).
$$

Yleisesti X:n r. momentti

$$
E(X^r) = (1-p) \cdot 0^r + p \cdot 1^r = p
$$

on tässä tapauksessa hyvin helppo laskea. Bernoullin jakauman $\mathrm{Ber}(p)$ momenttifunktio on

$$
\begin{aligned}
M(t) = E(\mathrm{e}^{tX}) &= P(X=0)\mathrm{e}^{t\cdot 0} + P(X=1)\mathrm{e}^{t\cdot 1} \\
&= (1-p) + p\mathrm{e}^t = 1 + p(\mathrm{e}^t - 1),
\end{aligned}
$$

joka on määritelty kaikilla $t \in \mathbb{R}$.

Esimerkki 4.3 (Sabharwal 1969). Olkoon n:n Bernoullin kokeen jonossa $X_1, X_2, \ldots, X_n$ onnistumistodennäköisyys $P(\mathrm{O}) = p$ ja vastaavasti $P(\mathrm{E}) = 1 - p$ (E = epäonnistuminen). Olkoon Y_n tapahtuman OE (osajono) esiintymisten lukumäärä koejonossa. Mikä on tällaisten osajonojen lukumäärän odotusarvo $E(Y_n)$? Määritellään ensin uusi satunnaismuuttuja

$$Z_i = h(X_i, X_{i+1}) = \begin{cases} 1, & \text{jos } X_i = \mathrm{O} \text{ ja } X_{i+1} = \mathrm{E}; \\ 0 & \text{muulloin,} \end{cases}$$

kun $i = 1, 2, \ldots, n-1$. Silloin

$$Y_n = \sum_{i=1}^{n-1} Z_i$$

ja

$$\begin{aligned} E\,Y_n &= \sum_{i=1}^{n-1} E(Z_i) \\ &= \sum_{i=1}^{n-1} p(1-p) = (n-1)p(1-p). \end{aligned}$$

Jos esimerkiksi $p = \frac{1}{2}$ ja $n = 101$, niin

$$E(Y_n) = \frac{n-1}{4} = 25.$$

□

Tehdään n riippumatonta Bernoullin koetta, joissa jokaisessa onnistumistodennäköisyys on p. Olkoon i. Bernoullin kokeen tulos satunnaismuuttuja X_i, joka saa arvon 1 tai 0. Silloin koesarjan tulos on riippumattomien samaa Bernoullin jakaumaa noudattavien satunnaismuuttujien jono $X_1, X_2, \ldots, X_n$, missä $P(X_i = 1) = p$ ja $P(X_i = 0) = q$, $i = 1, 2, \ldots, n$. Kun koe on tehty, tulos voisi olla esimerkiksi $111011000\ldots110$. Tällaisen tuloksen todennäköisyys (ennen koetta) olisi

$$ppp(1-p)p(1-p)(1-p)ppp\cdots pp(1-p) = p^k(1-p)^{n-k},$$

missä k on onnistumisten lukumäärä ja $n-k$ epäonnistumisten lukumäärä. Olkoon X onnistumisten lukumäärä n:ssä riippumattomassa Bernoullin kokeessa. Alaluvussa 3.6 totesimme, että X noudattaa binomijakaumaa parametrein n ja p. Silloin merkitään $X \sim \mathrm{Bin}(n, p)$. Binomijakauman todennäköisyysfunktio on

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n. \tag{4.2.2}$$

Esitetään nyt edellä mainittu binomijakauman luonnehdinta Bernoullin kokeiden avulla lauseen muodossa. Jatkossa oletetaan, että Bernoullin kokeet ovat toisistaan riippumattomat, vaikkei oletusta erikseen mainittaisikaan.

Lause 4.1 *Tehdään n riipumatonta Bernoullin koetta, joissa jokaisessa onnistumistodennäköisyys on p. Olkoon X onnistumisten lukumäärä. Silloin*

$$X \sim \mathrm{Bin}(n, p).$$

Todistus. Koska X on onnistumisten lukumäärä n:ssä riipumatomassa Bernoullin kokeessa, niin $X = X_1 + X_2 + \cdots + X_n$, missä $X_i \sim \mathrm{Ber}(p) = \mathrm{Bin}(1, p)$, $i = 1, 2, \ldots, n$ ovat riippumattomat ja noudattavat samaa Bernoullin jakaumaa. Merkitään nyt $X = S_n$ ja

$$S_n = X_1 + X_2 + \cdots + X_n = S_{n-1} + X_n.$$

Todistamme väitteen induktiolla.

Kun $n = 1$, niin oletuksen mukaan $X = X_1 \sim \mathrm{Ber}(p) = \mathrm{Bin}(1, p)$, joten väite pitää paikkansa tapauksessa $n = 1$. Teemme nyt induktio-oletuksen $S_{n-1} \sim \mathrm{Bin}(n-1, p)$ ja näytämme, että $S_n \sim \mathrm{Bin}(n, p)$.

Tapahtuma $\{S_{n-1} + X_n = k\}$ voidaan lausua yhdisteenä

$$\{S_{n-1} + X_n = k\} = \{S_{n-1} = k,\, X_n = 0\} \cup \{S_{n-1} = k-1,\, X_n = 1\},$$

missä $\{S_{n-1} = k,\, X_n = 0\}$ ja $\{S_{n-1} = k-1,\, X_n = 1\}$ ovat erillisiä tapahtumia. Silloin yhteenlaskusäännön nojalla

$$P(S_{n-1} + X_n = k) = P(S_{n-1} = k,\, X_n = 0) + P(S_{n-1} = k-1,\, X_n = 1).$$

Satunnaismuuttujat S_{n-1} ja X_n ovat oletuksen mukaan riippumattomat, joten

$$\begin{aligned}
&P(S_{n-1} + X_n = k) \\
&\quad = P(S_{n-1} = k)\, P(X_n = 0) + P(S_{n-1} = k-1)\, P(X_n = 1) \\
&\quad = \binom{n-1}{k} p^k (1-p)^{n-1-k} (1-p) + \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} p \\
&\quad = \binom{n-1}{k} p^k (1-p)^{n-k} + \binom{n-1}{k-1} p^k (1-p)^{n-k} \\
&\quad = \left[\binom{n-1}{k} + \binom{n-1}{k-1}\right] p^k (1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k},
\end{aligned}$$

missä viimeinen yhtäsuuruus seuraa siitä, että $\binom{n-1}{k} + \binom{n-1}{k-1} = \binom{n}{k}$ [Pascalin kolmio]. Näin on lause todistettu. $\square$

Esimerkki 4.4 Erään kasvin siementen itämistodennäköisyydeksi on ilmoitettu 0.8. Siemenen itäminen on tässä ”onnistuminen” ja itämistodennäköisyys on onnistumistodennäköisyys. Jos kylvetään 10 siementä ja siementen itämistapahtumat ovat toisistaan riippumattomat, niin kylvöä voidaan pitää kymmenenä riippumattomana Bernoullin kokeena, joissa onnistumistodennäköisyys on 0.8. Silloin itävien siementen lukumäärä $X \sim \mathrm{Bin}(10, 0.8)$, eli

$$f(x) = \binom{10}{x} 0.8^x \cdot 0.2^{10-x}, \qquad x = 0, 1, \ldots, 10.$$

Mikä on todennäköisyys, että vähemmän kuin 9 jyvää itää? Todennäköisyys

$$\begin{aligned} P(X < 9) = P(X \leq 8) &= 1 - \sum_{k=9}^{10} P(X = k) \\ &= 1 - 10 \cdot 0.8^9 \cdot 0.2 - 0.8^{10} = 0.6242. \end{aligned}$$

□

Laskemme usein muotoa $P(X \leq x)$ olevia todennäköisyyksiä, kuten edellisessä esimerkissä. Todennäköisyydet $P(X \leq x)$ määrittelevät jakauman kertymäfunktion

$$F(x) = P(X \leq x).$$

Kertymäfunktio määriteltiin alaluvussa 2.5.2. Binomijakauman kertymäfunktion arvot pisteissä $x = 0, 1, \ldots, n$ ovat

$$F(x) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}.$$

Lause 4.2 *Jos* $X \sim \mathrm{Bin}(n, p)$, *niin*

1. *X:n todennäköisyysfunktio $f(x)$ on*

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n$$

   kaikilla $n \in \mathbb{N}$ ja kaikilla $p \in [0, 1]$;

2. *X:n kertymäfunktio $F(y)$ on*

$$F(y) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \, \varepsilon(y - x)$$

   kaikilla $y \in \mathbb{R}$, missä $\varepsilon(y)$ on hyppyfunktio;

3. *X:n odotusarvo, varianssi ja momenttifunktio ovat*

$$\begin{gathered} \mu = E(X) = np, \qquad \mathrm{Var}(X) = np(1-p), \\ M(t) = E(\mathrm{e}^{tX}) = (1 - p + p\mathrm{e}^t)^n, \qquad -\infty < t < \infty. \end{gathered}$$

Todistus. 1. Binomijakauman todennäköisyysfunktio johdettiin Lauseen 4.1 todistuksessa.

2. Odotusarvo ja varianssi. Koska $X = X_1 + X_2 + \cdots + X_n$ on riippumattomien Bernoullin muuttujien $X_i \sim \mathrm{Ber}(p)$ summa, niin

$$\begin{aligned} E(X) &= E(X_1) + E(X_2) + \cdots + E(X_n) \\ &= p + p + \cdots + p = np \end{aligned}$$

ja

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) \\ &= p(1-p) + p(1-p) + \cdots + p(1-p) = np(1-p). \end{aligned}$$

3. Momenttifunktio on

$$\begin{aligned} M(t) &= E\big(\mathrm{e}^{tX}\big) \\ &= E\big(\mathrm{e}^{t(X_1+X_2+\cdots+X_n)}\big) = E\big(\mathrm{e}^{tX_1+tX_2+\cdots+tX_n}\big) \\ &= E\big(\mathrm{e}^{tX_1}\mathrm{e}^{tX_2}\cdots\mathrm{e}^{tX_n}\big) \\ &= E\big(\mathrm{e}^{tX_1}\big)\,E\big(\mathrm{e}^{tX_2}\big)\cdots E\big(\mathrm{e}^{tX_n}\big), \end{aligned}$$

missä viimeinen yhtäsuuruus seuraa lauseista 3.6 ja 3.10. Koska X_i ja X_j $(i \neq j)$ ovat riippumattomat, niin e^{tX_i} ja e^{tX_j} ovat riippumattomat (Lause 3.6) ja riippumattomien satunnaismuuttujien e^{tX_1}, e^{tX_2}, ..., e^{tX_n} tulon odotusarvo on yksittäisten tulon tekijöiden odotusarvojen tulo (Lause 3.10). Koska

$$M_{X_i}(t) = E(\mathrm{e}^{tX_i}) = 1 - p + p\mathrm{e}^t, \qquad i = 1, 2, \ldots, n,$$

niin

$$M(t) = (1 - p + p\mathrm{e}^t)^n \qquad \text{kaikilla } t \in \mathbb{R}.$$

Momenttifunktio itse asiassa määrittelee yksikäsitteisesti todennäköisyysfunktion (Lause 3.12). Näytämme kuitenkin vielä eksplisiittisesti, että binomitodennäköisyydet määrittelevät todennäköisyysfunktion. Koska Binomilauseen 2.6 perusteella

$$[p + (1-p)]^n = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = 1$$

kaikilla $p \in [0,1]$, niin todennäköisyydet $f(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$ määrittelevät todennäköisyysfunktion kaikilla $p \in [0,1]$ ja $n \geq 1$. Huomaa myös, että

$$M(0) = (1 - p + p\mathrm{e}^0)^n = [p + (1-p)]^n.$$

□

Seuraus 4.1 *Jos* $X_1 \sim \mathrm{Bin}(n_1, p)$ *ja* $X_2 \sim \mathrm{Bin}(n_2, p)$ *ovat riippumattomat, niin* $X_1 + X_2 \sim \mathrm{Bin}(n_1 + n_2, p)$.

Todistus. Koska Lauseen 4.2 mukaan X_1:n momenttifunktio on $(1-p+pe^t)^{n_1}$ ja X_2:n momenttifunktio on $(1-p+pe^t)^{n_2}$, niin satunnaismuuttujan X_1+X_2 momenttifunktio on Lauseen 3.13 mukaan $(1-p+pe^t)^{n_1+n_2}$. Mutta Lauseen 4.2 perusteella $(1-p+pe^t)^{n_1+n_2}$ on binomijakuman $\mathrm{Bin}(n_1+n_2,p)$ momenttifunktio. Tästä seuraa momenttifunktion yksikäsitteisyyden (Lause 3.12) nojalla, että $X_1+X_2\sim\mathrm{Bin}(n_1+n_2,p)$. □

Seurauslauseen 4.1 todistuksessa on käytetty esimerkin vuoksi yleistä momenttifunktiotekniikkaa. Tässä tapauksessa tulos saadaan kuitenkin helposti turvautumatta noin voimakkaisiin menetelmiin. Koska X_1 esittää onnistumisten lukumäärää n_1:ssä Bernoullin kokeessa ja X_2 onnistumisten lukumäärää n_2:ssa kokeessa, missä p on jokaisen kokeen onnistumistodennäköisyys, niin riippumattomien satunnaismuuttujien X_1 ja X_2 summa X_1+X_2 esittää onnistumisen lukumäärää (n_1+n_2):ssa kokeessa. Tämän perusteella saadaan tulos $X_1+X_2\sim\mathrm{Bin}(n_1+n_2,p)$. Analyyttisesti tulos voidaan tarkistaa laskemalla lauseke

$$\begin{aligned}
P(X_1+X_2=k) &= \sum_{i=0}^{n_1} P(X_1=i,\ X_2=k-i) \\
&= \sum_{i=0}^{n_1} P(X_1=i)\,P(X_2=k-i) \\
&= \sum_{i=0}^{n_1} \binom{n_1}{i} p^i(1-p)^{n_1-i}\binom{n_2}{k-i}p^{k-i}(1-p)^{n_2-k+i},
\end{aligned}$$

missä $\binom{n_2}{k-i}=0$ kaikilla $k-i>n_2$. Tästä seuraa

$$P(X_1+X_2=k)=p^k(1-p)^{n_1+n_2-k}\sum_{i=0}^{n_1}\binom{n_1}{i}\binom{n_2}{k-i}.$$

Soveltamalla hypergeometrista identiteettiä (ks. Lause 2.8)

$$\binom{n_1+n_2}{k}=\sum_{i=0}^{n_1}\binom{n_1}{i}\binom{n_2}{k-i}$$

saadaan kaivattu tulos.

4.3 Odotusaikojen jakaumat

Monissa sovelluksissa on kiinnostuksen kohteena odotusaika siihen hetkeen, että jokin tietty tapahtuma sattuu. Tässä alaluvussa käsitellään Bernoullin kokeisiin ja yksinkertaiseen satunnaisotantaan liittyviä odotusaikatehtäviä.

4.3.1 Odotusajat Bernoullin kokeissa

Tarkastellaan riippumattomien samaa Bernoullin jakaumaa noudattavien satunnaismuuttujien jonoa $X_1, X_2, \ldots, X_n$, missä $X_i \sim \mathrm{Ber}(p)$. Määritellään satunnaismuuttujat S_n ja W_r seuraavasti:

$$\begin{aligned} S_n &= X_1 + X_2 + \cdots + X_n, \\ W_r &= r\text{:ään onnistumiseen tarvittavien yritysten määrä.} \end{aligned}$$

Jos ajattelemme, että yhteen Bernoullin kokeeseen kuluu yhden yksikön pituinen aika, niin S_n vie n aikayksikköä. Nyt siis W_r on r:n onnistumisen saavuttamiseen tarvittava aika eli *odotusaika* ja sen mahdolliset arvot ovat r, $r+1$, $r+2$, ... Tiedämme, että $S_n \sim \mathrm{Bin}(n, p)$, mutta mikä on W_r:n jakauma?

Esimerkki 4.5 Heitetään harhatonta lanttia, kunnes saadaan kruunu (R). Olkoon W_1 tarvittavien heittojen lukumäärä. Tapahtuma $\{W_1 = x\}$ sattuu vain silloin, kun $(x-1)$:llä ensimmäisellä heitolla on saatu pelkkiä klaavoja (L) ja x. heitolla saadaan kruunu:

$$\underbrace{\mathrm{LLL}\ldots\mathrm{L}}_{x-1 \text{ kertaa}}\ \mathrm{R}.$$

Tästä seuraa, että

$$P(W_1 = x) = \frac{1}{2^x}, \qquad x = 1, 2, \ldots.$$

Satunnaismuuttujan W_1 odotusarvo on määritelmän mukaan

$$E(W_1) = \sum_{x=1}^{\infty} \frac{x}{2^x}. \tag{4.3.1}$$

Tiedämme, että

$$\sum_{x=0}^{\infty} p^x = 1 + p + p^2 + p^3 + \cdots = \frac{1}{1-p}, \qquad \text{kun } |p| < 1. \tag{4.3.2}$$

Kun derivoimme sarjan (4.3.2) termeittäin, saamme

$$0 + 1 + 2p + 3p^2 + \cdots = \sum_{x=0}^{\infty} (x+1) p^x = \frac{1}{(1-p)^2}, \qquad \text{kun } |p| < 1. \tag{4.3.3}$$

Koska sarjan (4.3.2) suppenemissäde on 1, suppenee derivointioperaation tuloksena saatu sarja (4.3.3) arvoilla $|p| < 1$. Sijoittamalla $p = \frac{1}{2}$ sarjaan (4.3.3) saadaan

$$\sum_{x=0}^{\infty} (x+1) \left(\frac{1}{2}\right)^x = 4,$$

joka voidaan esittää muodossa

$$\sum_{x=0}^{\infty} x\left(\frac{1}{2}\right)^x + \sum_{x=0}^{\infty}\left(\frac{1}{2}\right)^x = \sum_{x=0}^{\infty} x\left(\frac{1}{2}\right)^x + 2 = 4,$$

missä summa $\sum_{x=0}^{\infty}\left(\frac{1}{2}\right)^x = 2$ saadaan kaavasta (4.3.2). Nyt siis odotusarvo (4.3.1) on 2.

Jos kruunun todennäköisyys on p, niin silloin

$$P(W_1 = x) = \underbrace{(1-p)(1-p)\cdots(1-p)}_{x-1 \text{ kertaa}}\, p = (1-p)^{x-1}p$$

ja

$$\begin{aligned} E(W_1) &= \sum_{x=1}^{\infty} x(1-p)^{x-1}p = p\sum_{x=0}^{\infty}(x+1)(1-p)^x \\ &= p \cdot \frac{1}{[1-(1-p)]^2} = \frac{1}{p}, \end{aligned}$$

missä sarjan summa saadaan (4.3.3):n avulla. Satunnaismuuttuja W_1 on siis kruunun tai yleisemmin 'onnistumisen' odotusaika. Jakaumaa

$$P(W_1 = x) = (1-p)^{x-1}p, \qquad x = 1, 2, \ldots \tag{4.3.4}$$

kutsutaan *geometriseksi jakaumaksi*. Todennäköisyydet (4.3.4) todellakin määrittelevät jakauman, koska

$$\sum_{x=1}^{\infty} P(W_1 = x) = \sum_{x=1}^{\infty}(1-p)^{x-1}p = p \cdot \sum_{x=0}^{\infty}(1-p)^x = p \cdot \frac{1}{p} = 1.$$

□

Tapahtuma $\{W_r = x\}$ sattuu, kun $(x-1)$:ssä ensimmäisessä kokeessa on saatu $r-1$ onnistumista ja x. kokeessa saadaan onnistuminen:

$$\underbrace{\mathrm{OOEOE}\ldots\mathrm{E}}_{\substack{x-1 \text{ koetta,}\\ r-1 \text{ onnistumista,}\\ \text{kokeiden järjestys}\\ \text{mielivaltainen}}}\mathrm{O} \qquad \begin{cases} x.\ \text{koe,}\\ r.\ \text{onnistuminen}\end{cases}$$

Nyt siis $\{W_r = x\} = \{S_{x-1} = r-1,\ X_x = 1\}$. Koska X_i:t $(i = 1, 2, \ldots, x)$ ovat riippumattomat, niin myös S_{x-1} ja X_x ovat riippumattomat. Silloin

$$\begin{aligned} P(W_r = x) &= P(S_{x-1} = r-1)\, P(X_x = 1) \\ &= \binom{x-1}{r-1} p^{r-1}(1-p)^{x-r}p = \binom{x-1}{r-1} p^r(1-p)^{x-r}, \end{aligned} \tag{4.3.5}$$

koska $S_{x-1} \sim \mathrm{Bin}(x-1, p)$. Todennäköisyydet (4.3.5) määrittelevät ns. *negatiivisen binomijakauman.* Soveltamalla identiteettiä [ks. (2.4.5)]

$$\frac{r}{x}\binom{x}{r} = \binom{x-1}{r-1}$$

saadaan

$$P(W_r = x) = \frac{r}{x}\,P(S_x = r).$$

Toinen usein käyttökelpoinen identiteetti on

$$P(W_r > x) = P(S_x < r).$$

4.3.2 Geometrinen jakauma ja negatiivinen binomijakauma

Sanomme, että satunnaismuuttuja X noudattaa *negatiivista binomijakaumaa* parametrein r ja p, jos

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \qquad x = r, r+1, r+2, \ldots. \tag{4.3.6}$$

Merkitsemme silloin

$$X \sim \mathrm{NBin}(r, p).$$

Edellisessä pykälässä huomasimme, että odotusaika $W_r \sim \mathrm{NBin}(r, p)$. Kun $r = 1$, sanomme negatiivista binomijakaumaa *geometriseksi jakaumaksi.* Geometrisen jakauman todennäköisyysfunktio on siis

$$f(x) = p(1-p)^{x-1}, \qquad x = 1, 2, 3, \ldots. \tag{4.3.7}$$

Kun siis $X \sim \mathrm{NBin}(1, p)$, niin X:n noudattaa geometrista jakaumaa parametrilla p. Merkitsemme silloin $X \sim \mathrm{Geo}(p)$.

Lause 4.3 *Oletetaan, että* $X \sim \mathrm{NBin}(r, p)$.

1. *Funktio (4.3.6) on negatiivisen binomijakauman todennäköisyysfunktio kaikilla positiivisilla kokonaisluvuilla* r *ja kaikilla* $0 < p < 1$ *ja*

2.

$$E(X) = \frac{r}{p}, \qquad \mathrm{Var}(X) = \frac{r(1-p)}{p^2},$$

$$M(t) = E(\mathrm{e}^{tX}) = \frac{(p\mathrm{e}^t)^r}{[1-(1-p)\mathrm{e}^t]^r}, \qquad t < -\log(1-p).$$

Todistus. Johdamme ensin negatiivisen binomijakauman momenttifunktion suoraan määritelmän nojalla. Koska $M(t) = E(\mathrm{e}^{tX})$, niin momenttifunktio on

$$\begin{aligned}
E(\mathrm{e}^{tX}) &= \sum_{x=r}^{\infty} \mathrm{e}^{tx} \binom{x-1}{r-1} p^r (1-p)^{x-r} \\
&= p^r \sum_{y=0}^{\infty} \mathrm{e}^{t(y+r)} \binom{r+y-1}{r-1} p^r (1-p)^y \\
&= p^r \mathrm{e}^{tr} \sum_{y=0}^{\infty} \mathrm{e}^{ty} \binom{r+y-1}{y} (1-p)^y \\
&= p^r \mathrm{e}^{tr} \sum_{y=0}^{\infty} \mathrm{e}^{ty} (-1)^y \binom{-r}{y} (1-p)^y \\
&= p^r \mathrm{e}^{tr} \sum_{y=0}^{\infty} \binom{-r}{y} \left[-(1-p)\mathrm{e}^t\right]^y \\
&= p^r \mathrm{e}^{tr} \left[1-(1-p)\mathrm{e}^t\right]^{-r} = \left[\frac{p\mathrm{e}^t}{1-(1-p)\mathrm{e}^t}\right]^r.
\end{aligned}$$

Binomisarja $\sum_{y=0}^{\infty} \binom{-r}{y} \left[-(1-p)\mathrm{e}^t\right]^y$ suppenee (Lause 2.7), kun $(1-p)\mathrm{e}^t < 1$, joka on yhtäpitävä epäyhtälön $t < -\log(1-p)$ kanssa.

Koska $M(0) = 1$ kaikilla positiivisilla kokonaisluvuilla r ($r \in \mathbb{N}$) ja kaikilla $0 < p < 1$, niin (4.3.6) on todennäköisyysfunktio kaikilla $r \in \mathbb{N}$ ja kaikilla $0 < p < 1$. Odotusarvo ja varianssi saadaan laskemalla ensin $M(t)$:n 1. ja 2. derivaatta ja niiden avulla

$$E(X) = M'(0) \qquad \text{ja} \qquad \mathrm{Var}(X) = M''(0) - [M'(0)]^2.$$

□

Seuraus 4.2 *Jos $X \sim \mathrm{Geo}(p)$, niin $X \sim \mathrm{NBin}(1, p)$ ja*

1. *funktio (4.3.7) on geometrisen jakauman todennäköisyysfunktio kaikilla $0 < p < 1$ ja*

2.

$$E(X) = \frac{1}{p}, \qquad \mathrm{Var}(X) = \frac{1-p}{p^2},$$

$$M(t) = E(\mathrm{e}^{tX}) = \frac{p\mathrm{e}^t}{[1-(1-p)\mathrm{e}^t]}, \qquad t < -\log(1-p).$$

Olkoon Y epäonnistumisten lukumäärä Bernoullin toistokokeessa, ennen kuin saadaan r. onnistuminen. Koska r. onnistumiseen tarvittavien yritysten määrä $W_r \sim \mathrm{NBin}(r, p)$, niin

$$Y = W_r - r \qquad \text{ja} \qquad E(Y) = E(W_r) - r = \frac{r}{p} - r = \frac{r(1-p)}{p}.$$

Y:n varianssi on tietysti sama kuin W_r:n varianssi. Nyt siis $P(Y = y) = P(W_r = r + y)$ kaikilla $y = 0, 1, 2, \ldots$.

Nimitys ”negatiivinen binomijakauma” on peräisin esitystavasta

$$1 = p^r \cdot p^{-r} = p^r \left[1 - (1-p)\right]^{-r} = p^r \sum_{y=0}^{\infty} \binom{-r}{y} \left[-(1-p)\right]^y ,$$

mistä saadaan todennäköisyydet $P(W_r = y + r)$, $y = 0, 1, 2, \ldots$. Merkintä $\binom{-r}{y}$ on määritelmänsä mukaan

$$\binom{-r}{y} = \frac{(-r)^{(y)}}{y!} = (-1)^y \binom{r+y-1}{y},$$

missä $r > 0$ ja $y \geq 0$ ovat kokonaislukuja.

Esimerkki 4.6 Geometrisella jakaumalla ja negatiivisella binomijakaumalla on tärkeä merkitys esimerkiksi jonoteoriassa. Oletetaan, että joukko asiakkaita jonottaa pääsyä palvelutiskille. Olkoon todennäköisyys p, että jokaisella pienellä aikavälillä tulee 1 uusi asiakas (0 uutta asiakasta todennäköisyydellä $1 - p = q$). Silloin seuraavan asiakkaan odotusaika $W \sim \text{Geo}(p)$. Todennäköisyys $P(W > k)$, että seuraavan k:n aikayksikön aikana ei tule asiakasta, on

$$\begin{aligned} P(W > k) &= \sum_{j=k+1}^{\infty} q^{j-1} p = q^k (p + qp + q^2 p + \cdots) \\ &= q^k = 1 - P(W \leq k). \end{aligned}$$

□

Geometrisen jakauman kertymäfunktio on siis

$$\begin{aligned} F(k) = P(W \leq k) &= \sum_{i=1}^{k} (1-p)^{i-1} p \\ &= 1 - P(W > k) = 1 - q^k, \end{aligned}$$

missä $q = 1 - p$ ja $k = 1, 2, \ldots$. Geometrisen jakauman kertymäfunktion arvot saadaan geometrisesta sarjasta, josta jakauman nimi tulee.

Usein oletetaan, että myös asiakkaan palvelemiseen käytetty aika (palveluaika) noudattaa geometrista jakaumaa. Palveluajan jakaumalla on tietysti yleensä eri parametrin p arvo kuin palvelun odotusajan jakaumalla. Geometrisella jakaumalla on ”unohtamisominaisuus”, joka havaitaan laskemalla seuraava ehdollinen todennäköisyys:

$$P(W > k + s \mid W > k) = \frac{P(W > k + s)}{P(W > k)} = \frac{q^{k+s}}{q^k} = q^s. \tag{4.3.8}$$

Nyt siis todennäköisyys, että asiakkaan palveleminen kestää vielä s aikayksikköä, ei riipu siitä, kuinka kauan häntä on jo palveltu. Onneksi kuitenkin käytännössä palveluaika ei aina täysin noudata geometrista jakaumaa.

Esimerkki 4.7 Banachin tulitikkuongelma. Piippua polttelevalla matemaatikolla oli tapana pitää yksi tulitikkulaatikko oikeassa ja yksi vasemmassa taskussa. Joka kerta tikkua tarvitessaan hän valitsi taskun täysin satunnaisesti, joten kummankin taskun valintatodennäköisyys on $\frac{1}{2}$. Tarkastellaan tapahtumaa, että matemaatikko huomaa laatikon olevan tyhjä. Oletetaan, että kummassakin laatikossa oli alunperin N tikkua. Mikä on todennäköisyys, että toisessa laatikossa on täsmälleen k tikkua $(k = 0, 1, \ldots, N)$ silloin, kun matemaatikko havaitsee toisen laatikon olevan tyhjä?

Olkoon A tapahtuma, että matemaatikko huomaa oikeanpuoleisen laatikon olevan tyhjä ja samalla vasemman taskun laatikossa on k tikkua. Tapahtuma voi sattua täsmälleen silloin, kun oikeanpuoleisen taskun laatikosta valitaan tikku $(N+1)$. kerran ja yhteensä valintoja on tehty $N+1+N-k$ kappaletta. Teemme siis valintoja palauttamatta. Molemmissa laatikoissa on N tikkua, joten tapahtuma A on ekvivalentti tapahtuman $\{W_{N+1} = N+1+N-k\}$ kanssa. Saamme kaavalla (4.3.6) todennäköisyydeksi

$$P(W_{N+1} = N + 1 + N - k) = \binom{2N-k}{N}\left(\frac{1}{2}\right)^{2N-k+1}.$$

Koska myös todennäköisyys, että vasemmanpuoleinen laatikko huomataan tyhjäksi ja oikeanpuoleisessa on k tikkua, on $P(W_{N+1} = N+1+N-k)$, niin vastaus kysymykseen on

$$2\,P(W_{N+1} = N + 1 + N - k) = \binom{2N-k}{N}\left(\frac{1}{2}\right)^{2N-k}.$$

□

4.3.3 Odotusajat peräkkäisotannassa

Oletetaan, että populaatiossa on kahdenlaisia alkioita. Valitaan populaatiosta peräkkäisotos. Käytetään nyt apuna uurnamallia. Olkoon uurnassa *a valkoista* palloa ja *b mustaa* palloa eli yhteensä $a + b = N$ palloa. Poimitaan satunnaisvalinnalla palloja uurnasta yksitellen. Määritellään satunnaismuuttujat

$$\begin{aligned} S_n &= \text{valkoisten pallojen (onnistumisten) lukumäärä} \\ &\quad\; n\text{:ssä ensimmäisessä nostossa;} \\ W_r &= r\text{:n valkoisen pallon saamiseksi tarvittavien nostojen määrä.} \end{aligned}$$

Jos ajatellaan, että nostoon menee yksi aikayksikkö, niin W_r on r:n valkoisen pallon saamiseksi tarvittava odotusaika.

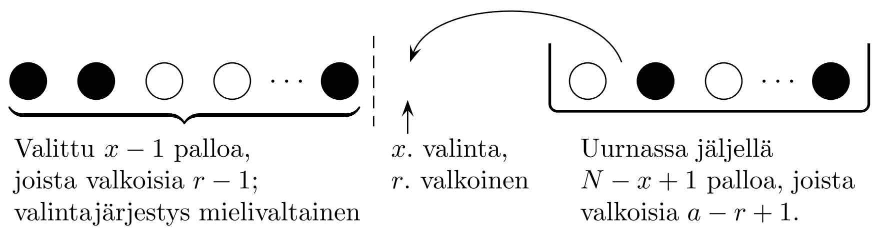
Jos otanta tehdään *palauttaen,* niin peräkkäiset nostot ovat riippumattomia Bernoullin kokeita, joissa onnistumistodennäköisyys on $p = a/N$. Tässä tapauksessa voidaan suoraan soveltaa edellä esitettyjä Bernoullin kokeita koskevia tuloksia.

Kun otanta tehdään *palauttamatta,* peräkkäiset nostot eivät ole riippumattomia, koska valkoisten pallojen suhteellinen osuus uurnassa riippuu siitä, mitä sieltä on jo valittu. Alaluvussa 2.6.1 osoitimme, että S_n noudattaa hypergeometrista jakaumaa, kun otanta tehdään palauttamatta (ks. myös alaluku 3.7.1). Silloin

$$P(S_n = x) = \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}, \tag{4.3.9}$$

kun $x = 0, 1, \ldots, n$. Mikä on todennäköisyys, että saamme x. nostossa r. valkoisen pallon?

Tapahtuma $\{W_r = x\}$ sattuu täsmälleen silloin, kun $x-1$ ensimmäisessä nostossa on saatu $r-1$ valkoista ja x. nostossa saadaan valkoinen:



Voimme siis kirjoittaa $\{W_r = x\} = \{S_{x-1} = r-1,\, X_x = 1\}$, missä $S_{x-1} \sim \mathrm{HGeo}(x-1, N, a/x)$ [ks. Esimerkki 3.11 ja (3.3.6)] ja $X_x = 1$, kun valitaan valkoinen pallo x. nostossa. Tästä seuraa, että

$$\begin{aligned} P(W_r = x) &= P(S_{x-1} = r-1, X_x = 1) \\ &= P(S_{x-1} = r-1)\, P(X_x = 1 \mid S_x = r-1) \\ &= \frac{\binom{a}{r-1}\binom{N-a}{x-r}}{\binom{N}{x-1}} \cdot \frac{a-r+1}{N-x+1}, \end{aligned} \tag{4.3.10}$$

kun $x = r, r+1, \ldots, N$.

Todennäköisyys (4.3.10) voidaan kirjoittaa lausekkeena

$$P(W_r = x) = \binom{x-1}{r-1} \frac{\binom{N-x}{a-r}}{\binom{N}{a}}, \tag{4.3.11}$$

joka on *negatiivisen hypergeometrisen jakauman* todennäköisyysfunktio. Koska $\binom{x-1}{r-1} = \frac{r}{x}\binom{x}{r}$, niin

$$P(W_r = x) = \frac{r}{x} \cdot \frac{\binom{x}{r}\binom{N-x}{a-r}}{\binom{N}{a}} = \frac{r}{n}\, P(S_x = r),$$

missä $S_x \sim \mathrm{HGeo}(x, N, a/N)$. Vastaavanlainen tulos saatiin otannassa palauttaen. Samoin on jälleen helppo nähdä, että

$$P(W_r > x) = P(S_x < r).$$

Merkitään $W_r \sim \mathrm{NHGeo}(r, N, p)$, missä $p = a/N$.

4.3.4 Hypergeometrinen jakauma ja negatiivinen hypergeometrinen jakauma

Olemme esitelleet hypergeometrisen jakauman tarkastelemalla otantaa palauttamatta (alaluku 2.6.1). Jakauman avulla voidaan siis ratkaista otantaan liittyviä todennäköisyystehtäviä. Hypergeometrisen jakauman momenttifunktiolla $M(t)$ ei ole olemassa siistiä lauseketta, vaikka se tietysti voidaan lausua määritelmänsä mukaan äärellisenä summana, koska satunnaismuuttujan arvojoukko on äärellinen. Hypergeometrisen jakauman odotusarvon ja varianssin laskeminen ei myöskään ole aivan helppo tehtävä.

Olemme merkinneet populaation alkioiden lukumäärää $N = a + b$, joista a kappaletta on tyyppiä A ja b kappaletta tyyppiä B. Esimerkiksi tuotepopulaatiossa on a viallista. Tyyppiä A olevien alkioden suhteellinen osuus on $p = a/N$. Tyyppiä A olevan alkion valinta on ”onnistuminen” ja tyypin B valinta ”epäonnistuminen”. Valitaan populaatiosta n:n alkion otos palauttamatta. Olkoon X onnistuneiden valintojen lukumäärä otoksessa. On selvää, että $0 \leq X \leq n$. Koska populaatiossa on pN kappaletta tyyppiä A olevia alkioita ja $(1-p)N$ kappaletta tyyppiä B, niin $X \leq pN$ ja $n-X \leq (1-p)N$. Siksi X:n arvoalue S on ehdon

$$\max\{0, n-(1-p)N\} \leq x \leq \min\{n, pN\}$$

toteuttavien kokonaislukujen x joukko.

Kun X noudattaa hypergeometrista jakaumaa $\mathrm{HGeo}(n, N, p)$, niin X:n todennäköisyysfunktio on

$$f(x) = P(X = x) = \frac{\binom{Np}{x}\binom{N-Np}{n-x}}{\binom{N}{n}}, \qquad x \in S. \tag{4.3.12}$$

Huomattakoon, että todennäköisyys (4.3.12) on määritelty myös arvoilla $x \notin S$, mutta silloin $f(x) = 0$.

Lause 4.4 *Oletetaan, että $X \sim \mathrm{HGeo}(n, N, p)$. Silloin*

$$E(X) = np \qquad \textit{ja} \qquad \mathrm{Var}(X) = \frac{N-n}{N-1}np(1-p).$$

Todistus. Hypergeometrisen jakauman odotusarvo laskettiin esimerkissä 3.11 ja alaluvussa 3.7.1. Varianssi voidaan laskea vastaavalla tavalla. □

Lause 4.5 *Oletetaan, että $Y \sim \mathrm{NHGeo}(r, N, p)$. Silloin*

$$E(Y) = r \cdot \frac{N+1}{Np+1} \qquad \textit{ja} \qquad \mathrm{Var}(Y) = \frac{rN(N+1)(1-p)(Np+1-r)}{(Np+1)^2(Np+2)}.$$

Mainitsimme jo alaluvussa 2.8.1, että binomijakaumaa voidaan käyttää hypergeometrisen jakauman likiarvona, kun N on suuri. Erityisesti, kun N on ääretön tai hyvin suuri (verrattuna otoskokoon), on yhdentekevää, käytetäänkö otantaa palauttaen vai palauttamatta. Oletetaan nyt, että

$$X_N \sim \mathrm{HGeo}(n, N, p) \qquad \text{ja} \qquad X \sim \mathrm{Bin}(n, p).$$

Kun parametrit n ja p ovat annettuja vakioita ja N kasvaa rajatta, voimme osoittaa, että X_N:n jakauma lähestyy X:n jakaumaa. Silloin siis

$$X_N \xrightarrow{d} X, \quad \text{kun} \quad N \to \infty.$$

Koska $X \sim \mathrm{Bin}(n, p)$, niin

$$X_N \xrightarrow{d} \mathrm{Bin}(n, p),$$

eli X_N:n jakauma lähestyy binomijakaumaa, jonka parametrit ovat n ja p. Sanomme myös, että X_N:n jakauma suppenee kohti X:n jakaumaa N:n kasvaessa. Kutsumme X:n jakaumaa X_N:n *asymptoottiseksi jakaumaksi.*

Lauseen 3.5 mukaan satunnaismuuttujilla on sama jakauma, jos niillä on sama kertymäfunktio. Voimme nyt tarkastella satunnaismuuttujien jonoa

$$\{\, X_N;\ N = 1, 2, \ldots \} = X_1, X_2, \ldots$$

ja vastaavaa kertymäfunktioiden jonoa

$$\{\, F_N;\ N = 1, 2, \ldots \} = F_1, F_2, \ldots,$$

missä $F_N(x)$ on X_N:n kertymäfunktio.

Määritelmä 4.2 Jono $\{\, X_N;\ N = 1, 2, \ldots \}$ suppenee jakaumaltaan kohti satunnaismuuttujaa X, jos

$$\lim_{N \to \infty} F_N(x) = F(x)$$

kaikissa pisteissä $x \in \mathbb{R}$, joissa X:n kertymäfunktio $F(x)$ on jatkuva.

Diskreettien satunnaismuuttujien tapauksessa voidaan helposti todistaa tulos, joka osoitaa, että suppenemista jakaumamielessä voidaan tarkastella yhtä hyvin myös todennäköisyysfunktioiden avulla.

Lause 4.6 *Olkoon* $\{\, X_N;\ N = 1, 2, \ldots \}$ *sellainen epänegatiivisten kokonaislukuarvoisten satunnaismuuttujien jono, että* X_N*:n todennäköisyysfunktio on* $f_N(k)$*,* $N = 1, 2, \ldots$*. Olkoon* X *epänegatiivinen kokonaislukuarvoinen satunnaismuuttuja, jonka todennäköisyysfunktio on* $f(k)$*. Silloin*

$$X_N \xrightarrow{d} X \Leftrightarrow \lim_{N \to \infty} f_N(k) = f(k)$$

kaikilla epänegatiivisilla kokonaisluvuilla k*.*

Todistus. Jätetään harjoitustehtäväksi. □

Lause 4.7 *Jos* $X_N \sim \mathrm{HGeo}(n, N, p)$, *niin*

$$X_N \xrightarrow{d} \mathrm{Bin}(n, p), \quad \textit{kun} \quad N \to \infty.$$

Todistus. Käytetään lausetta 4.6 ja osoitetaan, että $P(X_N = k) = f_N(k) \to f(k)$ kaikilla epänegatiivisilla kokonaisluvuilla k, kun $N \to \infty$. Yksityiskohdat jätetään lukijan pohdittavaksi. □

4.3.5 Tasajakauma

Diskreetti tasajakauma esiteltiin ensimmäisen kerran alaluvussa 2.5.4. Satunnaismuuttuja X, jonka arvoavaruus on $S = \{1, 2, \ldots, N\}$, noudattaa diskreettiä tasajakaumaa, jos

$$P(X = x) = \frac{1}{N}, \qquad k = 1, 2, \ldots, N.$$

Silloin merkitään $X \sim \mathrm{Tasd}(1, 2, \ldots, N)$, missä $N \geq 1$ on annettu positiivinen kokonaisluku. Jos $X \sim \mathrm{Tasd}(1, 2, \ldots, N)$, niin

$$E(X) = \frac{N+1}{2} \qquad \text{ja} \qquad \mathrm{Var}(X) = \frac{(N+1)(N-1)}{12}.$$

4.4 Poissonin jakauma

Satunnaismuuttuja X, jonka todennäköisyysfunktio on

$$f(x) = \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, \ldots \tag{4.4.1}$$

noudattaa *Poissonin jakaumaa* parametrilla $\lambda > 0$, joka on Poissonin jakauman odotusarvo. Silloin merkitään

$$X \sim \mathrm{Poi}(\lambda).$$

Poissonin jakaumalla on runsaasti sovelluksia eri aloilla. Sitä voidaan käyttää myös binomijakauman $\mathrm{Bin}(n, p)$ likiarvona, kun n on suuri ja p pieni. Silloin siis pätee

$$\binom{n}{x} p^x (1-p)^{n-x} \approx \frac{\mathrm{e}^{-np}(np)^x}{x!}.$$

Lause 4.8 *Olkoon* $X \sim \mathrm{Poi}(\lambda)$. *Silloin*

1. *funktio (4.4.1) on Poissonin jakauman todennäköisyysfunktio kaikilla* $\lambda > 0$ *ja*

2.

$$\mu = E(X) = \lambda, \qquad \mathrm{Var}(X) = \lambda,$$
$$M(t) = E(\mathrm{e}^{tX}) = \exp(\lambda \mathrm{e}^t - \lambda).$$

Todistus. Sovelletaan eksponenttifunktion sarjakehitelmää

$$\exp(\lambda) = \mathrm{e}^\lambda = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}. \tag{4.4.2}$$

1. Ensinnäkin $f(x) \geq 0$ kaikilla $x = 0, 1, 2, \ldots$, ja eksponenttifunktion sarjakehitelmän (4.4.2) perusteella

$$\sum_{x=0}^{\infty} f(x) = \sum_{x=0}^{\infty} \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!} = \mathrm{e}^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = \mathrm{e}^{-\lambda}\mathrm{e}^{\lambda} = 1.$$

2. Johdetaan ensin momenttifunktion $M(t)$ lauseke:

$$\begin{aligned} M(t) = E(\mathrm{e}^{tX}) &= \sum_{x=0}^{\infty} \mathrm{e}^{tx} \frac{\lambda^x}{x!} \mathrm{e}^{-\lambda} \\ &= \mathrm{e}^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda \mathrm{e}^t)^x}{x!} \\ &= \mathrm{e}^{-\lambda} \cdot \exp(\lambda \mathrm{e}^t) = \exp(\lambda \mathrm{e}^t - \lambda). \end{aligned}$$

Odotusarvo ja varianssi saadaan sitten laskemalla $M(t)$:n 1. ja 2. derivaatta ja soveltamalla identiteettejä

$$E(X) = M'(0) \qquad \text{ja} \qquad \mathrm{Var}(X) = M''(0) - [M'(0)]^2.$$

□

Riippumattomien Poissonin jakaumaa noudattavien satunnaismuuttujien summa noudattaa myös Poissonin jakaumaa.

Lause 4.9 *Olkoot $X_1, X_2, \ldots, X_n$ riippumattomat ja $X_i \sim \mathrm{Poi}(\lambda_i)$, $i = 1, 2, \ldots, n$. Olkoon $Y = X_1 + X_2 + \cdots + X_n$. Silloin*

$$Y \sim \mathrm{Poi}(\lambda),$$

missä $\lambda = \sum_{i=1}^{n} \lambda_i$.

Todistus. Seurauslauseen 3.1 mukaan

$$\begin{aligned} M_Y(t) &= \prod_{i=1}^{n} M_{X_i}(t) \\ &= \prod_{i=1}^{n} \exp(\lambda_i \mathrm{e}^t - \lambda_i) = \exp[(\mathrm{e}^t - 1)\lambda], \end{aligned}$$

missä $\lambda = \sum_{i=1}^{n} \lambda_i$. Lauseesta 3.12 seuraa sitten väite $Y \sim \mathrm{Poi}(\lambda)$. □

Jos riippumattomat X_1, X_2, $\ldots$, X_n noudattavat samaa Poissonin jakaumaa Poi(λ), niin Lauseen 4.9 mukaan niiden summa $Y = X_1 + X_2 + \cdots + X_n$ noudattaa Poissonin jakaumaa Poi($n\lambda$). Poissonin jakauma on hyvä binomijakauman Bin(n, p) likiarvo silloin, kun n on suuri ja p pieni.

Kun $X \sim \text{Bin}(n, p)$, niin binomitodennäköisyys on

$$f(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n. \tag{4.4.3}$$

Annetaan nyt p:n riippua n:stä ja merkitään lausekkeessa (4.4.3) $p = p_n$. Valitaan erityisesti

$$p_n = \frac{\lambda}{n}, \qquad n \geq 1.$$

Tarkastellaan nyt binomijakaumien jonoa

$$\text{Bin}(1, p_1),\ \text{Bin}(2, p_2),\ \text{Bin}(3, p_3), \ldots$$

ja vastaavaa satunnaismuuttujien X_1, X_2, X_3, $\ldots$ jonoa, missä $X_n \sim \text{Bin}(n, p_n)$, $n \geq 1$. Nyt siis

$$P(X_n = x) = \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}, \qquad 0 \leq x \leq n. \tag{4.4.4}$$

Merkitään todennäköisyyttä (4.4.4) lyhyesti $b_x(n)$

Kiinnitetään nyt x ja annetaan n:n kasvaa rajatta. Osoittautuu, että $b_x(n)$ suppenee kaikilla x. Valitaan ensin $x = 0$. Silloin saamme

$$\lim_{n\to\infty} b_0(n) = \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = \mathrm{e}^{-\lambda}. \tag{4.4.5}$$

Se on eräs keskeinen eksponenttifunktioon liittyvä kaava, joka pitäisi analyysin kurssien perusteella muistaa. Tulos (4.4.5) saadaan esimerkiksi Taylorin sarjan

$$\log(1-p) = -\sum_{n=1}^{\infty} \frac{p^n}{n}$$

avulla, kun sijoitetaan $p = \frac{\lambda}{n}$:

$$\begin{aligned} \log\left(1 - \frac{\lambda}{n}\right)^n &= n \log\left(1 - \frac{\lambda}{n}\right) = n\left(-\frac{\lambda}{n} - \frac{\lambda^2}{2n^2} - \frac{\lambda^3}{3n^3} - \cdots\right) \\ &= -\lambda - \frac{\lambda^2}{2n} - \frac{\lambda^3}{3n^2} - \cdots \\ &= -\lambda - \frac{1}{n}\left(\frac{\lambda^2}{2} + \frac{\lambda^3}{3n} + \cdots\right). \end{aligned} \tag{4.4.6}$$

Kun $n \to \infty$, niin $\frac{1}{n}\big(\frac{\lambda^2}{2} + \frac{\lambda^3}{3n} + \cdots\big) \to 0$ ja siksi $\log\big(1 - \frac{\lambda}{n}\big)^n \to -\lambda$.

Lasketaan seuraavaksi $b_x(n)$:n raja-arvo, kun $x > 0$. Tarkastellaan peräkkäisten binomitodennäköisyyksien suhdetta

$$\frac{b_{x+1}(n)}{b_x(n)} = \frac{n-x}{x+1}\left(\frac{\lambda}{n}\right)\left(1-\frac{\lambda}{n}\right)^{-1} = \frac{\lambda}{x+1}\left(\frac{n-x}{n}\right)\left(1-\frac{\lambda}{n}\right)^{-1},$$

missä $\frac{n-x}{n} \to 1$ ja $1 - \frac{\lambda}{n} \to 1$, kun $n \to \infty$. Tästä seuraa, että

$$\lim_{n\to\infty} \frac{b_{x+1}(n)}{b_x(n)} = \frac{\lambda}{x+1}. \tag{4.4.7}$$

Kun lähdetään tuloksesta (4.4.5) ja käytetään hyväksi raja-arvoa (4.4.7), saadaan

$$\begin{aligned}
\lim_{n\to\infty} b_1(n) &= \frac{\lambda}{1} \lim_{n\to\infty} b_0(n) = \lambda \mathrm{e}^{-\lambda}, \\
\lim_{n\to\infty} b_2(n) &= \frac{\lambda}{2} \lim_{n\to\infty} b_1(n) = \frac{\lambda^2}{1\cdot 2}\mathrm{e}^{-\lambda}, \\
&\vdots \\
\lim_{n\to\infty} b_x(n) &= \frac{\lambda}{x} \lim_{n\to\infty} b_{x-1}(n) = \frac{\lambda^x}{1\cdot 2\cdots x}\mathrm{e}^{-\lambda}.
\end{aligned}$$

Olemme siis näyttäneet, että

$$\lim_{n\to\infty} b_x(n) = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}, \tag{4.4.8}$$

missä raja-arvo on $P(X = x)$, kun $X \sim \mathrm{Poi}(\lambda)$. Tulos (4.4.8) tunnetaan *Poissonin raja-arvolakina.*

Satunnaismuuttujat noudattavat samaa jakaumaa, kun niillä on sama kertymäfunktio (Lause 3.5). Jos diskreetit satunnaismuuttujat noudattavat samaa jakaumaa, niin niillä on sama todennäköisyysfunktio. Jos satunnaismuuttujan X_n jakauma lähenee X:n jakaumaa n:n kasvaessa rajatta, niin X_n:n todennäköisyysfunktio lähenee X:n todennäköisyysfunktiota, mikäli jakaumat ovat diskreettejä (Lause 4.6). Vaikka edellä olemmekin johtaneet Poissonin raja-arvolain (4.4.8), esitetään tulos vielä *Poissonin lauseena.*

Lause 4.10 (Poissonin lause) *Olkoon* $X_n \sim \mathrm{Bin}(n, p)$. *Silloin*

$$X_n \xrightarrow{d} \mathrm{Poi}(\lambda),$$

kun $n \to \infty$ *siten, että* $np = \lambda$.

Todistus. Koska $np = \lambda$, voimme merkitä $p = \lambda/n$. Todistus perustuu

Lauseeseen 4.6. Jos $X_n \sim \mathrm{Bin}(n, p)$, niin

(4.4.9)
$$\begin{aligned} f_{X_n}(x) &= \binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^n \frac{n!}{(n-x)!\,n^x}\left(1-\frac{\lambda}{n}\right)^{-x} \\ &= \frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^n\left[\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\cdots\left(\frac{n-x+1}{n}\right)\right]\left(1-\frac{\lambda}{n}\right)^{-x}. \end{aligned}$$

Kiinteällä x:n arvolla

$$\lim_{n\to\infty}\left[\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\cdots\left(\frac{n-x+1}{n}\right)\right] = 1$$

ja

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-x} = 1.$$

Näistä tuloksista yhdessä raja-arvon (4.4.5) kanssa seuraa

$$\lim_{n\to\infty} f_{X_n}(x) = \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}.$$

Satunnaismuuttujan X_n jakauma lähestyy siis Poissonin jakaumaa $\mathrm{Poi}(\lambda)$, kun $n \to \infty$. $\square$

Poissonin jakaumaa sanotaan usein harvinaisten tapahtumien laiksi. Tämä luonnehdinta perustuu edellisessä lauseessa esitettyyn ominaisuuteen. Jos tehdään suuri määrä riippumattomia Bernoullin kokeita, joissa onnistumistodennäköisyys on hyvin pieni, niin silloin Lauseen 4.10 mukaan onnistumisten lukumäärä noudattaa likimain Poissonin jakaumaa. Esimerkiksi suuri määrä ihmisiä on päivittäin alttiina liikenneonnettomuuksille. Yksittäisen henkilön todennäköisyys (onnistumistodennäköisyys!) joutua onnettomuuteen on pieni, mutta onnettomuuksille alttiina olevien henkilöiden lukumäärä n on suuri. Silloin onnettomuuksien lukumäärä noudattaa likimain Poissonin jakaumaa.

Lause 4.11 *Olkoot X ja Y sellaiset riippumattomat satunnaismuuttujat, että $X \sim \mathrm{Poi}(\lambda_1)$ ja $Y \sim \mathrm{Poi}(\lambda_2)$. Silloin X:n ehdollinen jakauma ehdolla $X+Y$ on binomijakauma.*

Todistus. Olkoot m ja n sellaiset epänegatiiviset kokonaisluvut, että $m < n$.

Silloin

$$\begin{aligned}
P(X = m \mid X + Y = n) &= \frac{P(X = m, X + Y = n)}{P(X + Y = n)} \\
&= \frac{P(X = m, Y = n - m)}{P(X + Y = n)} \\
&= \frac{P(x = m)\, P(Y = n - m)}{P(X + Y = n)} \\
&= \frac{\mathrm{e}^{-\lambda_1}(\lambda_1^m/m!)\mathrm{e}^{-\lambda_2}[\lambda_2^{n-m}/(n-m)!]}{\mathrm{e}^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^n/n!} \\
&= \binom{n}{m}\frac{\lambda_1^m\lambda_2^{n-m}}{(\lambda_1+\lambda_2)^n} \\
&= \binom{n}{m}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^m\left(1-\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^{n-m}
\end{aligned}$$

on binomitodennäköisyys kaikilla $m = 0, 1, \dots, n$. Näin on lause todistettu.

□

Lauseella 4.11 on tärkeä merkitys esimekiksi frekvenssiaineistojen analyysissa.

Esimerkki 4.8 Tiedetään, että auto-onnettomuuksien lukumäärä aikayksikössä (esimerkiksi kuukaudessa) noudattaa Poissonin jakaumaa. Tarkastellaan eräällä tieosuudella lokakuussa sattuvien onnettomuuksien lukumäärää. Aikaisempien tilastojen perusteella voidaan olettaa, että auto-onnettomuuksien lukumäärä Z kyseisellä tieosuudella (kuukaudessa) noudattaa Poissonin jakaumaa $\mathrm{Poi}(\lambda)$. Onnettomuudet luokitellaan mahdollisten henkilövahinkojen mukaan vakaviin ja lieviin (jokainen onnettomuus kuuluu toiseen näistä luokista). Vakavien onnettomuuksien lukumäärä $X \sim \mathrm{Poi}(\lambda_1)$ ja lievien lukumäärä $Y \sim \mathrm{Poi}(\lambda_2)$. Lisäksi X ja Y ovat toisistaan riippumattomat. Koska $Z = X + Y$, niin $E(Z) = E(X) + E(Y)$ eli $\lambda = \lambda_1 + \lambda_2$.

Tutkijat valitsivat poliisin tiedostoista satunnaisesti valitun kuukauden (vuonna 2003) onnettomuudet. He havaitsivat onnettomuuksien lukumääräksi 120 ($n = 120$), mutta he eivät olleet vielä luokitelleet onnettomuuksia. Mitä jakaumaa noudattaa vakavien onnettomuuksien lukumäärä? Lauseen 4.11 perusteella

$$P(X = m \mid Z = 120) = \binom{120}{m}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^m\left(1-\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^{120-m},$$

$m = 0, 1, \dots, 120$. Vakavien onnettomuuksien lukumäärä noudattaa siis binomijakaumaa $\mathrm{Bin}\big(120, \frac{\lambda_1}{\lambda_1+\lambda_2}\big)$. Aikaisempien onnettomuustilastojen perusteella voimme arvioida parametrit λ_1 ja λ_2, joiden avulla saamme estimaatin parametrille $\frac{\lambda_1}{\lambda_1+\lambda_2}$. Kun tutkijat olivat luokitelleet nuo 120 onnettomuutta, aineistossa havaittiin 15 vakavaa onnettomuutta. Koska $E(X \mid Z = 120) = \lambda_1$, niin havainnon 15 pitäisi osua ”melko lähelle” arvoa λ_1. □

4.5 Poissonin prosessi

4.5.1 Laskuriprosessi

Stokastinen prosessi $\{ N(t),\ t \geq 0 \}$ on *laskuriprosessi,* jos $N(t)$ on ajankohtaan t mennessä sattuneiden ”tapahtumien” lukumäärä.

Esimerkki 4.9 Seuraavassa luetellaan esimerkkejä laskuririprosesseista.

1. Jos $N(t)$ on annetulla tieosuudella hetkeen t mennessä sattuneiden onnettomuuksien lukumäärä, niin $\{ N(t),\ t \geq 0 \}$ on tapahtumaan ”onnettomuus” liittyvä laskuriprosessi.

2. Olkoon $N(t)$ palvelutiskille tulleiden asiakkaiden lukumäärä hetkeen t mennessä. Tapahtuma on ”asiakkaan tulo palvelutiskille” ja $\{ N(t),\ t \geq 0 \}$ on tapahtumaan liittyvä laskuriprosessi.

3. $N(t)$ on vuoden alusta hetkeen t mennessä syntyneiden lasten lukumäärä kaupungissa A.

4. $N(t)$ on jalkapallojoukkueen A tekemien maalien lukumäärä kauden alusta ajankohtaan t mennessä.

□

Laskuriprosessin tulee toteuttaa seuraavat ominaisuudet:

1. $N(t) \geq 0$.

2. $N(t) \in \mathbb{N}$, eli $N(t)$ on kokonaislukuarvoinen.

3. Jos $s < t$, niin $N(s) \leq N(t)$.

4. Kun $s < t$, niin $N(t) - N(s)$ on välillä $(s, t]$ sattuneiden tapahtumien lukumäärä.

Laskuriprosessi on *riippumattomien lisäysten* prosessi, jos erillisillä aikaväleillä sattuvien tapahtumien lukumäärät ovat riippumattomat. Esimerkiksi satunnaismuuttujat $N(2)$ ja $N(10) - N(2)$ ovat riippumattomat, jos $N(t)$ on riippumattomien lisäysten laskuriprosessi. Laskuriprosessin *lisäykset ovat stationaariset,* jos millä tahansä välillä sattuvien tapahtumien lukumäärän jakauma riippuu vain välin pituudesta. Jos $N(t)$ on stationaarinen laskuriprosessi, niin satunnaismuuttujilla $N(t_2) - N(t_1)$ ja $N(t_2 + s) - N(t_1 + s)$ on sama jakauma kaikilla väleillä $(t_1, t_2]$ ja $(t_1 + s, t_2 + s]$, missä $t_2 > t_1$ ja $s > 0$.

4.5.2 Poissonin prosessin määritely

Poissonin prosessi on yksi tärkeimpiä laskuriprosesseja. Se määritellään seuraavasti:

Määritelmä 4.3 Laskuriprosessi $\{ N(t),\ t \geq 0 \}$ on Poissonin prosessi, jonka intensiteetti on λ ($\lambda > 0$), jos

1. $N(0) = 0$.
2. Prosessin lisäykset ovat riippumattomat.
3. Tapahtumien lukumäärä jokaisella h:n pituisella välillä noudattaa Poissonin jakaumaa, jonka odotusarvo on λh:

$$P[N(h+t) - N(t) = x] = \mathrm{e}^{-\lambda h} \frac{(\lambda h)^x}{x!}, \qquad x = 0, 1, \ldots$$

   kaikilla $h, t \geq 0$.

Laskuriprosessin osoittaminen Poissonin prosessiksi Määritelmän 4.3 avulla saattaa olla hankalaa. Ei ole mitään yksinkertaista keinoa tarkistaa esimerkiksi ehdon 3 pätevyyttä. Siksi esitetään vielä toinen määritelmä, jonka avulla voi olla helpompaa tunnistaa prosessi. Voidaan osoitaa, että määritelmät 4.3 ja 4.4 ovat yhtäpitävät.

Määritelmä 4.4 Laskuriprosessi $\{ N(t),\ t \geq 0 \}$ on Poissonin prosessi, jonka intensiteetti on λ ($\lambda > 0$), jos

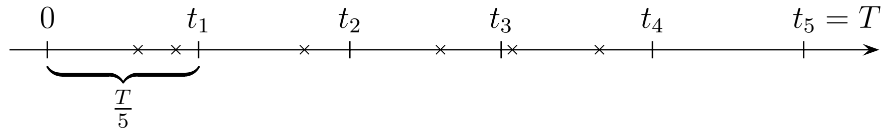
1. $N(0) = 0$.
2. Prosessin lisäykset ovat stationaariset ja riippumattomat.
3. $P\big(N(t+h) - N(t) = 1\big) = \lambda h + o(h)$.
4. $P\big(N(t+h) - N(t) \geq 2\big) = o(h)$.

Määritelmässä 4.4 käytetään merkintää $o(h)$. Sanomme, että funktio $f(\cdot) = o(h)$, jos

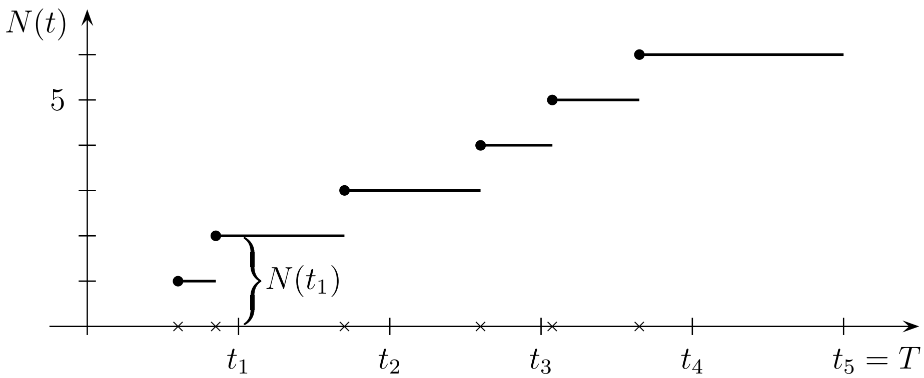
$$\lim_{h \to 0} \frac{f(h)}{h} = 0.$$

Esimerkki 4.10 Tieliikenneonnettomuudet. Havainnoidaan esimerkiksi jollain tieosuudella sattuvien auto-onnettomuuksien lukumäärää. Onnettomuuksien määrä noudattaa tavallisesti varsin hyvin Poissonin prosessia. $\square$

Tarkastellaan nyt hieman lähemmin Poissonin prosessin oletuksia. Oletetaan, että onnettomuuksien lukumäärä eräällä tieosuudella noudattaa aikavälillä $(0, T)$ Poissonin prosessia, jonka intensiteetti on λ. Aikaväli voi olla esimerkiksi ruuhka-aika tiettynä perjantai-iltapäivänä klo 15–19 ja tieosuus jokin ulosmenotie. Oheisessa kuviossa on havaitut onnettomuudet merkitty aika-akselille.

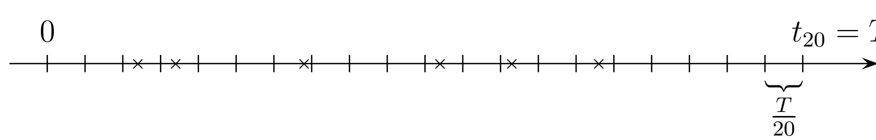


Tarkasteluväli $(0, T]$ on jaettu viiteen yhtä pitkään osaväliin, joiden pituudet ovat $T/5$. Nyt esimerkiksi 1. osavälillä sattuneiden onnettomuuksien lukumäärä on $N(t_1) - N(0) = N(t_1)$, joka on siis hetkeen t mennessä sattuneiden onnettomuuksien lukumäärä. Kuvioon 4.2 on piirretty prosessin $\{ N(t),\ t \in (0, T] \}$ realisaatio, missä havaintoina ovat kyseiset onnettomuudet.



Kuvio 4.2. Poissonin prosessin $\{ N(t),\ t \in (0, T] \}$ erään realisaation kuvaaja.

Määritelmän 4.4 oletuksen 2 mukaan lisäykset $N(t_1) - N(0)$, $N(t_2) - N(t_1)$, $N(t_3) - N(t_2)$, $N(t_4) - N(t_3)$ ja $N(t_5) - N(t_4)$ ovat riippumattomat ja noudattavat samaa jakaumaa. Määritelmän 4.4 oletukset 3 ja 4 tarkoittavat, että tapahtumat (onnettomuudet) sattuvat yksittäin ja samalla intensiteetillä koko tarkastelujakson ajan. Koska tapahtumat ovat erillisiä pisteitä, niin aina voidaan valita niin hienojakoinen välin ositus, että kullakin osavälillä on korkeintaan 1 tapahtuma. Jos tarkastelemassamme esimerkkitapauksessa valitaan osavälin pituudeksi $T/20$, sattuu tässä osituksessa kullekin osavälille korkeintaan 1 tapahtuma. Riippuen tietysti kulloisestakin havaintojaksosta, kuinka hienojakoinen ositus tarvitaan.



Todennäköisyys, että T/n:n pituiselle osavälille sattuu havainto, on Määritelmän 4.4 oletuksen 3 mukaan

$$P\left[N\left(t + \frac{T}{n}\right) - N(t) = 1\right] = \lambda \cdot \frac{T}{n} + o\left(\frac{T}{n}\right).$$

Vastaavasti todennäköisyys, että osavälillä sattuu enemmän kuin yksi havainto, on häviävän pieni, sillä Määritelmän 4.4 oletuksen 4 mukaan

$$P\left[N\left(t+\frac{T}{n}\right)-N(t)\geq 2\right]=o\left(\frac{T}{n}\right).$$

Voimme siis olettaa, että kullakin osavälillä sattuu vain 0 tai 1 tapahtumaa, kun n on riittävän suuri.

Määritellään nyt satunnaismuuttujat

$$X_i=N\left(\frac{iT}{n}\right)-N\left(\frac{(i-1)T}{n}\right),\qquad i=1,2,\dots,n.$$

Muuttujia X_i voidaan käsitellä toisistaan riippumattomina Bernoullin jakaumaa noudattavina satunnaismuuttujina:

$$X_i\sim \mathrm{Ber}\left(\frac{\lambda T}{n}\right),\qquad i=1,2,\dots,n.$$

Koko välillä $(0,T]$ havaittujen tapahtumien lukumäärä on

$$S_n=X_1+X_2+\cdots+X_n,$$

joka noudattaa binomijakaumaa $\mathrm{Bin}\big(n,\frac{\lambda T}{n}\big)$. Koska $E(S_n)=n\cdot\frac{\lambda T}{n}=\lambda T$ kaikilla $n\in\mathbb{N}$, niin $E(S_n)=\lambda T$, kun $n\to\infty$. Voimme siis soveltaa Poissonin lausetta (Lause 4.10), jonka mukaan S_n noudattaa Poissonin jakaumaa $\mathrm{Poi}(\lambda T)$, kun n kasvaa rajatta. Näin esimerkiksi todennäköisyys, että välillä $(0,T]$ sattuu x onnettomuutta, on

$$P\big(N(T)=x\big)=\frac{\mathrm{e}^{-\lambda T}(\lambda T)^x}{x!}.$$

Todennäköisyys riippuu vain välin pituudesta T ja intensiteetistä $\lambda>0$.

4.5.3 Satunnaistapahtumat tila-avaruudessa

Poissonin prosessilla mallinnetaan myös ilmiöitä, jotka tapahtuvat satunnaisesti tila-avaruudessa. Silloin Määritelmän 4.4 ehdot voidaan luonnehtia seuraavasti:

1. *Riippumattomuus.* Erillisillä alueilla sattuvien tapahtumien lukumäärät ovat riippumattomat.

2. *Yksittäisyys.* Todennäköisyys, että alueella sattuu enemmän kuin yksi tahtuma, on häviävän pieni.

3. *Homogeenisuus.* Tapahtumat sattuvat samalla intensiteetillä koko tarkasteltavalla alueella.

Tarkastellaan esimerkiksi Poissonin prosessia tasossa. Silloin todennäköisyys, että pinta-alaltaan A:n kokoisella alueella sattuu x tapahtumaa, on

$$f_A(x) = \frac{\mathrm{e}^{-\lambda A}(\lambda A)^x}{x!}, \qquad x = 0, 1, \ldots,$$

missä λ on tapahtumien lukumäärän odotusarvo yhtä pinta-alayksikköä kohti. Jos Poissonin prosessia noudattavat tapahtumat sattuvat kolmiulotteisessa avaruudessa, niin silloin V:n kokoiseen tilaan osuu x tapahtumaa todennäköisyydellä

$$f_V(x) = \frac{\mathrm{e}^{-\lambda V}(\lambda V)^x}{x!}, \qquad x = 0, 1, \ldots,$$

missä λ on tapahtumien lukumäärän odotusarvo yhtä tilavuus-yksikköä kohti.

Esimerkki 4.11 Leipomo valmistaa suuren erän pullataikinaa, josta tehdään rusinapullia. Leipuri haluaa, että ainakin 95 % pullista sisältää vähintään 2 rusinaa. Kuinka monta rusinaa pullaa kohti pitäisi sekoittaa taikinaan?

Olkoon pullan tilavuus $V = 1$. Kun rusinat sekoitetaan hyvin taikinaan, on kaikilla pullilla sama todennäköisyys sisältää rusinoita (homogeenisuus). Koska taikina on suuri, ovat eri pulliin sattuvien rusinoiden lukumäärät toisistaan riippumattomat. Todennäköisyys, että pieneen pullaan sattuu enemmän kuin yksi rusina, on hyvin pieni.

Tässä tilanteessa on kyse Poissonin prosessista 3-ulotteisessa tila-avaruudessa. Pullassa on x rusinaa todennäköisyydellä

$$f(x) = \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, 2, \ldots$$

ja ainakin 2 rusinaa todennäköisyydellä

$$\begin{aligned} P(X \geq 2) &= 1 - P(X < 2) \\ &= 1 - P(X = 0) - P(X = 1) \\ &= 1 - \mathrm{e}^{-\lambda} - \mathrm{e}^{-\lambda}\lambda. \end{aligned}$$

Leipuri vaatii, että

$$1 - \mathrm{e}^{-\lambda} - \mathrm{e}^{-\lambda}\lambda \geq 0.95.$$

Epäyhtälö toteutuu, kun $\lambda \geq 4.74$, joten rusinoita on sekoitettava taikinaan 5 rusinaa pullaa kohti. □

4.6 Kaksiulotteiset jakaumat

Tilastollisissa sovelluksissa tarkastellaan tavallisesti useita muuttujia samanaikaisesti. Esimerkiksi haastattelututkimuksessa valitaan opiskelijoista satunnaisotosotos. Jokaiselta otokseen osuneelta kysytään useita kysymyksiä ja lisäksi saadaan haastateltavien taustatiedot kuten ikä, sukupuoli, asuinpaikka jne. Otosavaruudessa on siis määritelty useita muuttujia (kysymykset ja taustamuuttujat). Tällainen asetelma mahdollistaa muuttujien välisten riippuvuuksien tarkastelun. Seuraavassa esitellään usean muuttujan jakaumiin liittyvää käsitteistöä. Ensin käsitellään kahden muuttujan tapaus yksityiskohtaisesti. Sen jälkeen on suoraviivaista yleistää tarkastelu usean muuttujan tapaukseen.

Määritelmä 4.5 Olkoot X ja Y samassa otosavaruudessa määritellyt diskreetit satunnaismuuttujat ja olkoon kaksiulotteisen diskreetin satunnaismuuttujan (X, Y) arvoavaruus S. Tapahtuman "$\{X = x\}$ ja $\{Y = y\}$ sattuvat" todennäköisyyttä merkitään $P(X = x, Y = y) = f(x, y)$. Funktio $f(x, y)$ on (X, Y):n todennäköisyysfunktio (tnf), jolla on seuraavat ominaisuudet:

1. $0 \leq f(x, y) \leq 1$,
2. $\sum_{(x,y) \in S} f(x, y) = 1$ ja
3. $P[(X, Y) \in A] = \sum_{(x,y) \in A} f(x, y)$, missä $A \subset S$.

Funktiota $f(x, y)$ sanotaan myös X:n ja Y:n yhteisjakauman todennäköisyysfunktioksi. Moniulotteista satunnaismuuttujaa kutsutaan satunnaisvektoriksi (SV).

Esimerkki 4.12 Olkoon (X, Y) satunnaisvektori, jonka arvoavaruus on

$$S = \{(0, 1), (0, 2), (1, 0), (1, 1), (2, 0)\}$$

ja todennäköisyysfunktio

$$f(x, y) = c(x + 2y), \qquad (x, y) \in S.$$

Todennäköisyysfunktion ominaisuuksista seuraa, että

$$\sum_{(x,y) \in S} c(x + 2y) = c(2 + 4 + 1 + 3 + 2) = 12c = 1,$$

joten $c = \frac{1}{12}$. Silloin esimerkiksi

$$P(X > Y) = f(1, 0) + f(2, 0) = \frac{3}{12}$$

ja

$$P(X \geq Y) = f(1, 0) + f(2, 0) + f(1, 1) = \frac{6}{12}.$$

□

4.6.1 Reunajakauma ja ehdollinen jakauma

Jos (X, Y) on kaksiulotteinen satunnaisvektori, niin X ja Y ovat satunnaismuuttujia. Satunnaismuuttujan X *reunajakauman todennäköisyysfunktio,* jota merkitään $f_X(x)$, on on X:n todennäköisyysfunktio, kun Y:tä ei oteta huomioon. Satunnaismuuttujan X *ehdollisen jakauman todennäköisyysfunktio,* jota merkitään $f_1(x \mid y)$, on on X:n todennäköisyysfunktio, kun Y:n arvo $Y = y$ on kiinnitetty.

Määritelmä 4.6 Olkoon diskreetin satunnaisvektorin (X, Y) todennäköisyysfunktio $f(x, y)$ ja arvoavaruus S. Silloin satunnaismuuttujat X ja Y ovat diskreettejä ja niiden reunajakaumien todennäköisyysfunktiot ovat

$$f_X(x) = \sum_{y \in S_Y} f(x, y), \quad x \in S_X; \qquad f_Y(y) = \sum_{x \in S_X} f(x, y), \quad y \in S_Y,$$

missä S_X on X:n ja S_Y Y:n arvoavaruus. Satunnaismuuttujat *X ja Y ovat riippumattomat* jos ja vain jos

$$P(X = x,\, Y = y) = P(X = x)\, P(Y = y) \tag{4.6.1}$$

kaikilla $x \in S_X$ ja $y \in S_Y$. Jos X ja Y eivät ole riippumattomia, niin ne ovat riippuvia. Todennäköisyysfunktion avulla ehto (4.6.1) voidaan lausua muodossa:

$$f(x, y) = f_X(x) f_Y(y) \qquad \text{kaikilla } x \in S_X \text{ ja } y \in S_Y.$$

Merkitään $A = \{X = x\}$ ja $B = \{Y = y\}$, missä $(x, y) \in S$. Silloin $A \cap B = \{X = x,\, Y = y\}$. Koska

$$P(A \cap X) = P(X = x, Y = y) = f(x, y)$$

ja

$$P(B) = P(Y = y) = f_Y(y) > 0 \qquad (\text{koska } y \in S_Y),$$

niin

$$P(A \mid B) = \frac{P(A \cap X)}{P(B)} = \frac{f(x, y)}{f_Y(y)}.$$

Siksi voimme määritellä ehdollisen todennäköisyysfunktion seuraavasti:

Määritelmä 4.7 Jos diskreetin satunnaisvektorin (X, Y) todennäköisyysfunktio on $f(x, y)$ ja arvoavaruus S, niin *X:n ehdollinen todennäköisyysfunktio ehdolla $Y = y$* on

$$f_1(x \mid y) = \frac{f(x, y)}{f_Y(y)}, \qquad (x, y) \in S$$

ja *Y:n ehdollinen todennäköisyysfunktio ehdolla $X = x$* on

$$f_2(y \mid x) = \frac{f(x, y)}{f_X(x)}, \qquad (x, y) \in S.$$

Esimerkki 4.13 Esimerkissä 4.12 käsitellyn satunnaisvektorin (X, Y) todennäköisyysfunktio on

$$f(x, y) = \frac{x + 2y}{12}, \qquad \text{kun } (x, y) \in S,$$

missä $S = \{(0,1), (0,2), (1,0), (1,1), (2,0)\}$. Huomaa, että $f(x, y) = 0$, kun $(x, y) \notin S$. X:n ja Y:n reunajakaumien todennäköisyysfunktiot ovat

$$f_X(x) = \sum_{y=0}^{2} f(x, y) \qquad \text{ja} \qquad f_Y(y) = \sum_{x=0}^{2} f(x, y).$$

X:n ehdollinen todennäköisyysfunktio ehdolla $Y = y$ on

$$f(x \mid y) = \frac{f(x, y)}{f_Y(y)}, \qquad x = 0, 1, 2.$$

Saadaan siis kolme X:n ehdollista todennäköisyysfunktiota:

$$\begin{aligned} f_1(x \mid 0) &= \frac{x}{3}, && x = 0, 1, 2; \\ f_1(x \mid 1) &= \frac{x + 2}{5}, && x = 0, 1; \\ f_1(x \mid 2) &= 1, && x = 0. \end{aligned}$$

Vastaavalla tavalla saadaan kolme Y:n ehdollista todennäköisyysfunktiota ehdolla $X = x$. Satunnaismuuttujat X ja Y eivät ole riippumattomat, koska esimerkiksi

$$f(0, 1) = \frac{1}{6} \neq f_X(0) f_Y(1) = \frac{1}{2} \cdot \frac{5}{12} = \frac{5}{24}.$$

X:n ja Y:n riippuvuutta (vs. riippumattomomuutta) on luontevaa tarkastella ehdollisen jakauman avulla. Y:n ehdollinen todennäköisyysfunktio ehdolla $X = 1$ on

$$f_2(y \mid 1) = \frac{f(1, y)}{f_X(1)} = \frac{1 + 2y}{12} \Big/ \frac{1}{3} = \frac{1 + 2y}{4}, \qquad y = 0, 1.$$

Koska

$$f_2(y \mid 1) \neq f_Y(y),$$

voimme jälleen päätellä, että X ja Y eivät ole riippumattomat. □

Ehdollisen jakauman odotusarvoa kutsutaan jakauman *ehdolliseksi odotusarvoksi.* X:n ehdollinen odotusarvo ehdolla $Y = y$ on

$$E(X \mid Y = y) = \sum_{x \in S_X} x f(x \mid y) \tag{4.6.2}$$

ja Y:n ehdollinen odotusarvo ehdolla $X = x$ on

$$E(Y \mid X = x) = \sum_{y \in S_Y} y f(y \mid x). \tag{4.6.3}$$

Näitä odotusarvoja merkitään myös

$$E(X \mid y) = \mu_{X|y} \qquad \text{ja} \qquad E(Y \mid x) = \mu_{Y|x}.$$

Vastaavasti määritellään *X:n ehdollinen varianssi* ehdolla $Y = y$ ja Y:n ehdollinen varianssi ehdolla $X = x$. Y:n ehdollinen varianssi ehdolla $X = x$ on

$$\begin{aligned} \operatorname{Var}(Y \mid x) &= E\big[(Y - \mu_{Y|x})^2 \mid x\big] \\ &= \sum_{y \in S_Y} (y - \mu_{Y|x})^2 f(y \mid x), \end{aligned} \tag{4.6.4}$$

jota merkitään myös $\operatorname{Var}(Y \mid x) = \sigma^2_{Y|x}$. Samalla periaatteella voidaan ehdollisen jakauman avulla määritellä mikä tahansa jakauman ehdollinen tunnusluku, kuten esimerkiksi ehdolliset momentit tai ehdollinen mediaani.

Kun $E(Y \mid x)$ lasketaan eri x:n arvoilla, riippuu tulos yleensä x:n arvosta. Jos halutaan korostaa $E(Y \mid x)$:n riippuvuutta x:stä, merkitään esimerkiksi $E(Y \mid x) = g(x)$. Silloin ehdollinen odotusarvo määrittelee funktion $g(x)$.

Esimerkki 4.14 Esimerkissä 4.13 määritettiin ehdolliset todennäköisyysfunktiot $f_1(x \mid 0), f_1(x \mid 1)$ ja $f_1(x \mid 1)$, kun (X, Y):n yhteisjakauman todennäköisyysfunktio on

$$f(x, y) = \frac{x + 2y}{12}, \qquad (x, y) \in S.$$

Lasketaan nyt ehdolliset odotusarvot $E(X \mid y)$, $y = 0, 1, 2$:

$$\begin{aligned} E(X \mid 0) &= 0 + 1 \cdot \frac{1}{3} + 2 \cdot \frac{2}{3} = \frac{5}{3}, \\ E(X \mid 1) &= 0 \cdot \frac{2}{5} + 1 \cdot \frac{3}{5} + 2 \cdot 0 = \frac{3}{5}, \\ E(X \mid 2) &= 0 \cdot 1 + 1 \cdot 0 + 2 \cdot 0 = 0. \end{aligned}$$

Ehdolliset varianssit $\operatorname{Var}(X \mid y)$, $y = 0, 1, 2$, ovat vastaavasti:

$$\begin{aligned} \operatorname{Var}(X \mid 0) &= \left(1 - \frac{5}{3}\right)^2 \cdot \frac{1}{3} + \left(2 - \frac{5}{3}\right)^2 \cdot \frac{2}{3} = \frac{2}{9}, \\ \operatorname{Var}(X \mid 1) &= \left(0 - \frac{3}{5}\right)^2 \cdot \frac{2}{2} + \left(1 - \frac{3}{5}\right)^2 \cdot \frac{3}{5} + \left(2 - \frac{3}{5}\right)^2 \cdot 0 = \frac{6}{25}, \\ \operatorname{Var}(X \mid 2) &= (0 - 0)^2 \cdot 1 + (1 - 0)^2 \cdot 0 + (2 - 0)^2 \cdot 0 = 0. \end{aligned}$$

□

4.6.2 Satunnaismuuttujien funktion jakauma

Usein tarvitaan satunnaismuuttujien X ja Y jonkin funktion $h(X,Y)$ jakaumaa. Funktio $h(X,Y)$ voi olla esimerkiksi muotoa $X+Y$, XY, X^2+Y^2 jne. Jos h on jokin reaaliarvoinen funktio $h(x,y)$, voimme määritellä uuden satunnaismuuttujan $Z = h(X,Y)$. Olkoon satunnaisvektorin (X,Y) arvoalue S. Merkitään

$$A_z = \{\,(x,y) \in S \mid h(x,y) = z\,\}.$$

Silloin todennäköisyys $P(Z=z)$ voidaan laskea seuraavasti:

$$P(Z=z) = \sum_{(x,y)\in A_z} f(x,y).$$

Lasketaan siis yhteen todennäköisyydet $f(x,y)$ kaikkissa pisteissa (x,y), jotka toteuttavat ehdon $h(x,y)=z$. Tällä tavalla voidaan johtaa Z:n todennäköisyysfunktio.

Esimerkki 4.15 Oletetaan, että satunnaisvektorin (X,Y) (Esimerkki 4.12) todennäköisyysfunktio on

$$f(x,y) = \frac{x+2y}{12}, \qquad (x,y) \in S,$$

missä $S = \{(0,1),(0,2),(1,0),(1,1),(2,0)\}$. Määritellään kokonaislukuarvoinen satunnaismuuttuja

$$Z = h(X,Y) = XY.$$

Silloin Z:n arvojen joukko on $S_z = \{0,1\}$, ja vastaavasti

$$\begin{aligned} A_1 &= \{\,(x,y) \mid xy = 1\,\} = \{(1,1)\}, \\ A_0 &= \{\,(x,y) \mid xy = 0\,\} = \{(0,1),(0,2),(1,0),(2,0)\}. \end{aligned}$$

Nyt siis

$$\begin{aligned} P(Z=0) &= P\big((X,Y)\in A_0\big) = \frac{3}{4}, \\ P(Z=1) &= P\big((X,Y)\in A_1\big) = \frac{1}{4}, \end{aligned}$$

joten $Z \sim \mathrm{Ber}\big(\frac{1}{4}\big)$. □

4.6.3 Ehdollinen odotusarvo

Ehdollinen odotusarvo esiteltiin jo alaluvussa 3.3.2 (identiteetti 3.3.7). Alaluvussa 4.6.1 ehdollinen odotusarvo luonnehdittiin ehdollisen jakauman odotusarvona (ks. (4.6.2) ja (4.6.4)). Määritelmän mukaan

$$E(X \mid Y=y) = \sum_x x f(x \mid y) = \sum_x x \frac{f(x,y)}{f_Y(y)}.$$

Huomaa, että $E(X \mid Y = y)$ on y:n funktio, eli $E(X \mid Y = y) = h(y)$. Merkitään $E(X \mid Y) = h(Y)$, missä siis $E(X \mid Y)$ on satunnaismuuttuja, joka saa arvoja $E(X \mid Y = y)$, $y \in S_Y$. Voimme nyt laskea satunnaismuuttujan $E(X \mid Y)$ odotusarvon, joka on $E(X)$. Monissa sovelluksissa odotusarvon laskeminen on luontevinta ehdollistamisen kautta.

Lause 4.12 *Olkoot X ja Y mitkä tahansa kaksi satunnaismuuttujaa, joilla on odotusarvo. Silloin $E[E(X \mid Y)] = E(X)$.*

Todistus. Odotusarvon määritelmän mukaan

$$\begin{aligned} E[E(X \mid Y)] &= \sum_y E(X \mid Y = y) f_Y(y) \\ &= \sum_y \sum_x x \frac{f(x,y)}{f_Y(y)} f_Y(y) \\ &= \sum_y \sum_x x f(x,y) \\ &= \sum_x x \sum_y f(x,y) \\ &= \sum_x x f_X(x) = E(X), \end{aligned}$$

missä $\sum_y f(x,y) = f_X(x)$ on X:n reunajakauma. □

Ehdollinen varianssi

$$\begin{aligned} Var(X \mid Y = y) &= E[(X - E(X \mid Y = y))^2 \mid Y = y] \\ &= E(X^2 \mid Y = y) - [E(X \mid Y = y)]^2 \end{aligned}$$

määriteltiin alaluvussa 4.6.1 (ks. identiteetti (4.6.4)). Ehdollinen varianssi $Var(X \mid Y)$ on satunnaismuuttuja, joka saa arvoja $Var(X \mid Y = y)$. Koska

$$Var(X \mid Y) = E(X^2 \mid Y) - [E(X \mid Y)]^2,$$

niin

$$\begin{aligned} E[Var(X \mid Y)] &= E[E(X^2 \mid Y)] - E[E(X \mid Y)]^2 \\ &= E(X^2) - E[E(X \mid Y)]^2. \end{aligned} \tag{4.6.5}$$

Lauseen 4.12 mukaan $E[E(X \mid Y)] = E(X)$ ja $E[E(X^2 \mid Y)] = E(X^2)$, joten

$$Var[E(X \mid Y)] = E[E(X \mid Y)]^2 - [E(X)]^2. \tag{4.6.6}$$

Laskemalla yhtälöt (4.6.5) ja (4.6.6) puolittain yhteen, saadaan seuraavassa lauseessa esitettävä tulos.

Lause 4.13 *Mille tahansa satunnaismuutujille X ja Y pitää paikkansa identiteetti*

$$Var(X) = E[Var(X \mid Y)] + Var[E(X \mid Y)],$$

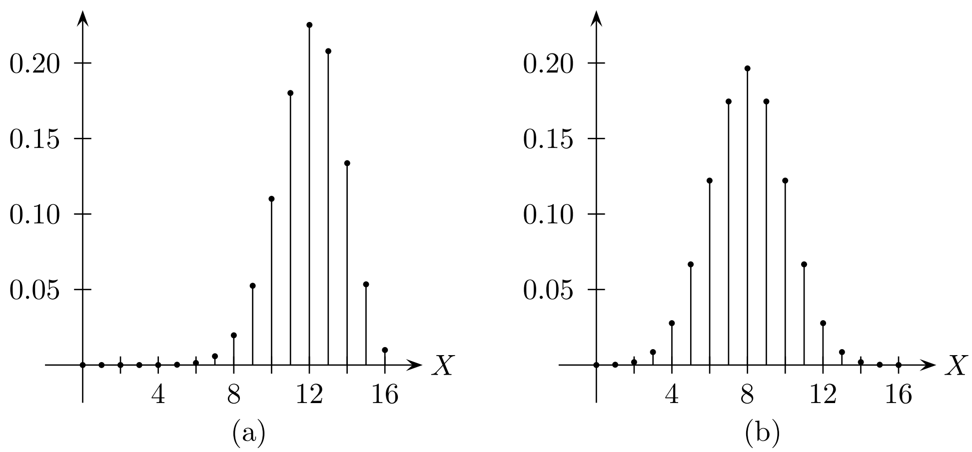
jos odotusarvot ovat olemassa.

4.6.4 Symmetrinen jakauma

Symmetriaan perustuvaa argumentointia voidaan usein hyödyntää todennäköisyyksien laskemisessa.

Symmetria pisteen suhteen. Jos $P(X = b + x) = P(X = b - x)$ kaikilla x, niin X:n jakauma on *symmetrinen pisteen b suhteen*. Satunnaismuuttuja X on symmetrinen b:n suhteen jos ja vain jos $X - b$ on symmetrinen origon suhteen. Silloin

$$P(X \leq b - x) = P(X \geq b + x).$$



Kuvio 4.3. Binomijakauman kuvaajat, kun (a) $X \sim \mathrm{Bin}(16, 0.75)$ (b) $X \sim \mathrm{Bin}(16, 0.50)$.

Esimerkiksi binomijakauma $\mathrm{Bin}(16, 0.50)$ on symmetrinen pisteen 8 suhteen, mutta binomijakauma $\mathrm{Bin}(16, 0.75)$ ei ole symmetrinen (Kuvio 4.3). Binomijakaumassa $\mathrm{Bin}(16, 0.50)$ on

$$P(X = 8 + x) = P(X = 8 - x)$$

kaikilla x. Silloin jokaista $a \in \mathbb{R}$ kohti

$$P(X \leq 8 - a) = P(X \geq 8 + a).$$

Esimerkki 4.16 Hatussa on 3 korttia, jotka on numeroitu yhdestä kolmeen. Valitaan hatusta peräkkäin satunnaisesti palauttamatta 2 korttia. Olkoon X ensiksi valitun kortin numero ja Y toisen kortin numero. Selvästikin

$$P(X = i) = f_X(i) = \frac{1}{3}, \qquad i = 1, 2, 3.$$

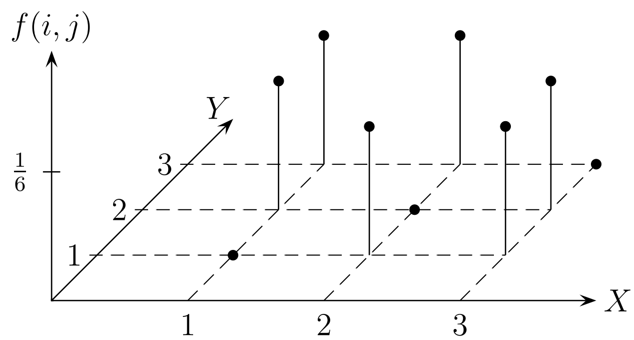
On helppo havaita, että toisen valinnan tulos Y riippuu 1. valinnan tuloksesta:

$$P(Y=1 \mid X=1)=0, \qquad P(Y=i \mid X=1)=\frac{1}{2}, \quad i=2,3;$$
$$P(Y=2 \mid X=2)=0, \qquad P(Y=i \mid X=2)=\frac{1}{2}, \quad i=1,3;$$
$$P(Y=3 \mid X=3)=0, \qquad P(Y=i \mid X=3)=\frac{1}{2}, \quad i=1,2.$$

Koska $P(X=i, Y=j) = P(X=i)\,P(Y=j \mid X=i)$, niin X:n ja Y:n yhteisjakauman todennäköisyysfunktio on

$$f(i,j) = f_X(i) f_2(j \mid i) = \begin{cases} \frac{1}{6}, & \text{kun } i \neq j; \\ 0, & \text{kun } i = j;\ 1 \leq i \leq 3,\ 1 \leq j \leq 3. \end{cases} \tag{4.6.7}$$

Satunnaisvektorin (X,Y) arvojoukko $S = \{(1,2),(1,3),(2,1),(2,3),(3,1),$



Kuvio 4.4. Satunnaisvektorin (X,Y) yhteisjakauman todennäköisyysfunktio $f(i,j)$, kun X on 1. valinta ja Y on 2. valinta palauttamatta joukosta $\{1,2,3\}$.

$(3,2)\}$, sillä $P(X=i, Y=j) > 0$ kaikilla $(i,j) \in S$ ja $P(X=i, Y=j) = 0$, jos $(i,j) \notin S$. $\square$

Jakauma (4.6.7) on esimerkki symmetrisestä 2-ulotteisesta jakaumasta. Diskreetin satunnaisvektorin (X,Y) *jakauma on symmetrinen,* jos sen todennäköisyysfunktio $f(x,y)$ on *symetrinen funktio.* Se tarkoittaa sitä, että

$$f(x,y) = f(y,x) \qquad \text{kaikilla } (x,y) \in S,$$

missä S on (X,Y):n arvoalue.

4.6.5 Kaksiulotteinen Bernoullin jakauma

Bernoullin jakaumaa noudattava satunnaismuuttuja $X \sim \mathrm{Ber}(p)$ on eräs yksinkertaisimpia ajateltavissa olevia satunnaismuuttujia. Sen todennäköisyysfunktio on

$$f_X(x) = p^x (1-p)^{1-x}, \qquad x \in \{0,1\},$$

kun $0 \leq p \leq 1$. Bernoullin jakauma on binomijakauman erikoistapaus siten, että $X \sim \mathrm{Bin}(1, p)$.

Kaksiulotteista Bernoullin jakaumaa noudattava satunnaismuuttuja (X, Y) voi saada arvot $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$. Sen todennäköisyysfunktio on

$$f(x, y) = p_{xy}, \qquad x \in \{0,1\} \quad \text{ja} \quad y \in \{0,1\}, \tag{4.6.8}$$

missä $p_{00} + p_{01} + p_{10} + p_{11} = 1$. Todennäköisyysfunktio voidaan esittää myös muodossa

$$f(x, y) = p_{00}^{(1-x)(1-y)} p_{01}^{(1-x)y} p_{10}^{x(1-y)} p_{11}^{xy},$$

kun $x \in \{0,1\}$ ja $y \in \{0,1\}$; muualla $f(x, y) = 0$. Todennäköisyydet $P(X = x, Y = y) = p_{xy}$ on esitetty Taulukossa 4.1

Taulukko 4.1. Kaksiulotteisen Bernoullin jakauman todennäköisyysfunktio.

$f(x,y)$	$y=0$	$y=1$	$f_X(x)$
$x=0$	p_{00}	p_{01}	$1-p_1$
$x=1$	p_{10}	p_{11}	p_1
$f_Y(y)$	$1-p_2$	p_2	1

"Reunatodennäköisyydet" määritellään $p_{00}+p_{01} = p_1$ ja $p_{00}+p_{10} = p_2$. On helppo havaita, että $X \sim \mathrm{Ber}(p_1)$ ja $Y \sim \mathrm{Ber}(p_2)$. Näiden reunajakaumien todennäköisyysfunktiot ovat siis

$$f_X(x) = p_1^x (1 - p_1)^{1-x}, \qquad x \in \{0,1\}$$

ja

$$f_Y(y) = p_2^y (1 - p_2)^{1-y}, \qquad y \in \{0,1\}.$$

Nyt esimerkiksi Y:n ehdollinen todennäköisyysfunktio ehdolla $X = 1$ on

$$f_2(y \mid 1) = \frac{p_{1y}}{p_1}, \qquad y \in \{0,1\} \tag{4.6.9}$$

kun $p_1 > 0$. Satunnaismuuttujan Y ehdollinen jakauma ehdolla $X = 1$ on siis $\mathrm{Ber}(p_{11}/p_1)$. Satunnaismuuttujat X ja Y ovat riippumattomat täsmälleen silloin, kun $P(X = x, Y = y) = P(X = x)\, P(Y = y)$ kaikilla $x \in \{0,1\}$ ja $y \in \{0,1\}$.

Koska $p_{00} + p_{01} + p_{10} + p_{11} = 1$, niin kaksiulotteinen Bernoullin jakauma voidaan luonnehtia kolmella parametrilla. Jakauman kolme "luonnollista" parametria ovat

$$\begin{aligned} p_1 &= E(X) = P(X = 1), \\ p_2 &= E(Y) = P(Y = 1), \\ p_{11} &= E(XY) = P(X = 1, Y = 1). \end{aligned}$$

Kun (X, Y) noudattaa kaksiulotteista Bernoullin jakaumaa parametrein p_1, p_2 ja p_{11}, niin merkitään $(X, Y) \sim \mathrm{Ber}(p_1, p_2, p_{11})$.

4.7 Satunnaismuuttujien funktion odotusarvo

Olkoot X ja Y diskreetit satunnaismuuttujat, joiden yhteinen todennäköisyysfunktio $f(x,y)$ on on määritelty arvoavaruudessa S. Olkoon $h(X,Y)$ satunnaismuuttujien X ja Y reaaliarvoinen funktio. Silloin

$$E[h(X,Y)] = \sum_{(x,y)\in S} h(x,y)f(x,y)$$

on satunnaismuuttujan $h(X,Y)$ odotusarvo, mikäli summa on olemassa.

Huomaa, että odotusarvon $E[h(X,Y)]$ olemassaolo tarkoittaa sitä, että summa

$$\sum_{(x,y)\in S} h(x,y)f(x,y)$$

suppenee itseisesti eli summa

$$\sum_{(x,y)\in S} |h(x,y)|f(x,y)$$

suppenee ja on äärellinen. Tästä seuraa, että $E[h(X,Y)]$ on olemassa. Funktio $V = h(X,Y)$ on satunnaismuuttuja, jonka todennäköisyysfunktio $g(v)$ on määritelty arvoavaruudessa $S_v = \{\, v \mid v = h(x,y),\ (x,y)\in S \,\}$. Silloin

$$E[h(X,Y)] = E(V) = \sum_{v\in S_v} vg(v).$$

4.7.1 Momentit

Monilla odotusarvoilla on omat nimensä, koska niillä on tärkeä rooli jakaumateoriassa. Olkoot X_1 ja X_2 diskreetit satunnaismuuttujat, joiden yhteisjakauman todennäköisyysfunktio $f(x_1,x_2)$ on on määritelty arvoavaruudessa S. Olkoon $h(X_1,X_2)$ satunnaismuuttujien X_1 ja X_2 reaaliarvoinen funktio. Määritellään esimerkiksi seuraavat odotusarvot:

1. Jos $h(X_1,X_2) = X_i$, niin

   $$E[h(X_1,X_2)] = E(X_i) = \mu_i$$

   on X_i:n *odotusarvo,* $i = 1,2$.

2. Jos $h(X_1,X_2) = (X_i-\mu_i)^2$, niin

   $$E[h(X_1,X_2)] = E[(X_i-\mu_i)^2] = \sigma_i^2$$

   on X_i:n *varianssi,* $i = 1,2$.

3. Jos $h(X_1, X_2) = (X_1 - \mu_1)(X_2 - \mu_2)$, niin

$$E[h(X_1, X_2)] = E[(X_1 - \mu_1)(X_2 - \mu_2)] = \sigma_{12}$$

   on X_1:n ja X_2:n *kovarianssi.*

Odotusarvo μ_i ja varianssi σ_i^2 voidaan laskea joko yhteisjakauman todennäköisyysfunktion $f(x_1, x_2)$ tai reunajakauman todennäköisyysfunktion $f_i(x_i)$ avulla.

Vastaavalla tavalla voidaan määritellä kaikkien kertalukujen momentit: Olkoon r positiivinen kokonaisluku.

1. Jos $h(X_1, X_2) = X_i^r$, niin

$$E[h(X_1, X_2)] = E(X_i^r)$$

   on X_i:n r. *momentti,* $i = 1, 2$.

2. Jos $h(X_1, X_2) = (X_i - \mu_i)^r$, niin

$$E[h(X_1, X_2)] = E[(X_i - \mu_i)^r]$$

   on X_i:n r. *keskusmomentti,* $i = 1, 2$.

3. Jos $h(X_1, X_2) = X_1^r X_2^s$, niin

$$E[h(X_1, X_2)] = E(X_1^r X_2^s)$$

   on X_1:n ja X_2:n kertalukua $r + s$ oleva *yhteismomentti.*

Esimerkiksi kovarianssin laskemisessa tarvitaan X_1:n ja X_2:n yhteismomentti $E(X_1 X_2)$.

4.7.2 Satunnaisvektorin momenttifunktio

Satunnaisvektorin (X, Y) yhteisjakauman momenttifunktio on

$$\begin{aligned} M(t, s) &= E[\exp(tX + sY)] \\ &= \sum_{x_i \in S_X} \sum_{y_j \in S_Y} \exp(tx_i + sy_j) f(x_i, y_j), \end{aligned}$$

mikäli odotusarvo on olemassa nollan ympäristössä. Silloin on siis olemassa sellainen positiiviluku $a > 0$, että odotusarvo $E[\exp(tX + sY)]$ on olemassa kaikilla $(t, s) \in \{\, (t, s) \mid t^2 + s^2 < a \,\}$ jollain $a > 0$. Edellä on käytetty merkintää $\exp(tX + sY) = \mathrm{e}^{tX + sY}$.

4.8 Riippumattomat satunnaismuuttujat

Riippumattomuuden määritelmän mukaan tapahtumat $\{X = x\}$ ja $\{Y = y\}$ ovat riippumattomat jos ja vain jos $P(X = x, Y = y) = P(X = x)\,P(Y = y)$ eli

$$f(x, y) = f_X(x) f_Y(y), \tag{4.8.1}$$

missä $f(x, y)$ on X:n ja Y:n yhteisjakauman todennäköisyysfunktio, $f_X(x)$ on X:n ja $f_Y(y)$ Y:n todennäköisyysfunktio. Satunnaismuuttujat *X ja Y ovat riippumattomat* jos ja vain jos yhtäsuuruus (4.8.1) pitää paikkansa kaikilla $x \in S_X$ ja $y \in S_Y$, missä S_X on X:n ja S_Y on Y:n arvojoukko. Voidaan helposti osoittaa, että X ja Y ovat riippumattomat, jos ja vain jos

$$F(x, y) = F_X(x) F_Y(y) \tag{4.8.2}$$

kaikilla $x \in S_X$ ja $y \in S_Y$, missä $F_X(x)$ on X:n ja $F_Y(y)$ Y:n kertymäfunktio (reunakertymäfunktio).

Satunnaismuuttujien X ja Y ovat riippumattomuus voidaan luonnehtia myös ehdollisten jakaumien avulla. Jos Määritelmässä 4.7

$$f(y \mid x) = f_Y(y)$$

kaikilla $x \in S_X$ ja $y \in S_Y$, kun $f_X(x) \neq 0$, niin X ja Y ovat riippumattomat. Tämä tarkoittaa sitä, että tieto X:n arvosta ei vaikuta Y:n todennäköisyyteen. Vastaavasti pitää paikkansa, että X ja Y ovat riippumattomat jos ja vain jos X:n ehdollinen todennäköisyysfunktio ehdolla $Y = y$ ei riipu y:stä.

Olkoot $Y_1, Y_2, \ldots, Y_n$ jossain otosavaruudessa määritellyt diskreetit satunnaismuuttujat. Muuttujien $Y_1, Y_2, \ldots, Y_n$ *yhteisjakauman todennäköisyysfunktio* on

$$f(y_1, y_2, \ldots, y_n) = P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n).$$

Muuttujat $Y_1, Y_2, \ldots, Y_n$ ovat riippumattomat, jos

$$f(y_1, y_2, \ldots, y_n) = f_1(y_1) f_2(y_2) \cdots f_n(y_n) \tag{4.8.3}$$

kaikilla $y_i \in S_i$, $i = 1, 2, \ldots, n$, missä $f_i(y_i)$ on Y_i:n todennäköisyysfunktio ja S_i on Y_i:n arvoavaruus.

4.8.1 Riippumattomat kokeet

Olkoot $\mathcal{E}_1$ ja $\mathcal{E}_2$ riippumattomat satunnaiskokeet. Oletetaan, että satunnaismuuttujan X arvo määräytyy vain satunnaiskokeen $\mathcal{E}_1$ tuloksen ja Y:n arvo vain satunnaiskokeen $\mathcal{E}_2$ tuloksen perusteella. Silloin tapahtumat $\{X = x\}$ ja $\{Y = y\}$ määräytyvät eri satunnaiskokeista, jotka ovat riippumattomat (katso alaluku 3.6 ja määritelmä (3.6.1)). Siksi tapahtumat $\{X = x\}$ ja $\{Y = y\}$

ovat riippumattomat. Koska riippumattomuus pätee kaikilla mahdollisilla x:n ja y:n arvoilla, niin X ja Y ovat riippumattomat. *Jos satunnaismuuttujien arvot määräytyvät eri satunnaiskokeista, jotka ovat riippumattomat, niin satunnaismuuttujat ovat riippumattomat.*

Tehdään esimerkiksi Bernoullin toistokoe, jossa on $r + s$ toistoa ja onnistumistodennäköisyys p. Olkoon X onnistumisten lukumäärä r:ssä ensimmäisessä kokeessa ja Y onnistumisten lukumäärä s:ssä viimeisessä kokeessa. Koska X ja Y riippuvat eri kokeista, jotka ovat riippumattomat, niin X ja Y ovat riippumattomat. Silloin (4.8.1):n mukaan

$$f(x,y) = f_X(x)f_Y(y) = \binom{r}{x}\binom{s}{y}p^{x+y}(1-p)^{r+s-x-y},$$

missä $x = 0, 1, \ldots, r$ ja $y = 0, 1, \ldots, s$.

4.8.2 Samoin jakautuneet riippumattomat (SJR) satunnaismuuttujat

Riippumattomia satunnaismuuttujia Y_1, Y_2, $\ldots$, Y_n, joista jokainen noudattaa samaa jakaumaa, sanotaan samoin jakautuneiksi riippumattomiksi (sjr) satunnaismuuttujiksi. Silloin puhutaan usein lyhyesti sjr satunnaismuuttujista. Vastaava englanninkielinen termi on iid (independent, identically distributed).

Jos esimerkiksi $Y_i \sim \mathrm{Poi}(\lambda)$, $i = 1, 2, \ldots, n$, niin

$$f_1(y) = f_2(y) = \cdots = f_n(y) = \frac{\lambda^y \mathrm{e}^{-\lambda}}{y!}.$$

Silloin (4.8.3):n nojalla

$$\begin{aligned} f(y_1, y_2, \ldots, y_n) &= \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \frac{\lambda^{y_i}\mathrm{e}^{-\lambda}}{y_i!} \\ &= \frac{1}{y_1!\,y_2!\ldots y_n!}\lambda^y \mathrm{e}^{-n\lambda}, \end{aligned}$$

missä $y = \sum_{i=1}^{n} y_i$.

4.8.3 Riippumattomien satunnaismuuttujien funktio

Olkoot X ja Y riippumattomat satunnaismuuttujat. Silloin

$$f(x,y) = f_X(x)f_Y(y) \qquad \text{kaikilla } x, y \in S,$$

missä S on satunnaisvektorin (X, Y) arvoavaruus. Määritellään nyt satunnaismuuttujat $U = g(X)$ ja $V = h(Y)$, missä $g(\cdot)$ riippuu vain X:stä ja $h(\cdot)$ vain Y:stä. Silloin Lauseen 3.6 mukaan U ja V ovat riippumattomat.

Väite todistettiin tarkastelemalla U:n ja V:n mielivaltaisten arvojen u ja v todennäköisyyttä. Olkoon $A_u = \{\, x \in S_X \mid g(x) = u \,\}$ ja $B_v = \{\, y \in S_Y \mid g(y) = v \,\}$. Koska kaikilla U:n ja V:n arvoilla u ja v

$$
\begin{aligned}
(4.8.4) \qquad & P(U = u, V = v) \\
&= P(X \in A_u, Y \in B_v) \\
&= P(X \in A_u)\, P(Y \in B_v) \qquad (X \text{ ja } Y \text{ riippumattomat}) \\
&= P(U = u)\, P(V = v),
\end{aligned}
$$

niin U ja V ovat riippumattomat. Kun sijoitetaan identiteettiin (4.8.4)

$$P(U = u, V = v) = \sum_{x \in A_u} \sum_{y \in B_v} f(x, y),$$

$$P(U = u) = \sum_{x \in A_u} f_X(x) \qquad \text{ja} \qquad P(V = v) = \sum_{y \in B_v} f_Y(y),$$

niin saadaan

$$\sum_{x \in A_u} \sum_{y \in B_v} f(x, y) = \Bigl(\sum_{x \in A_u} f_X(x) \Bigr) \Bigl(\sum_{y \in B_v} f_Y(y) \Bigr).$$

4.9 Multinomijakauma ja moniulotteinen hypergeometrinen jakauma

Binomijakauma ja multinomijakauma ovat keskeisen tärkeitä tilastollisissa sovelluksissa, koska niitä tarvitaan esimerkiksi riippumattomien koetoistojen tulosten frekvenssijakaumien käsittelyssä. Alaluvussa 3.6 esitettiin, miten binomijakauma saadaan Bernoullin kokeiden avulla. Kun toistetaan kokeita, joissa on useampia kuin kaksi tulosvaihtoehtoa, tulosten frekvenssijakauma voidaan kuvata multinomijakauman avulla. Laajennetaan ensin binomijakauma *trinomijakaumaksi.*

Tarkastellaan koetta, jossa on kolme toisensa poissulkevaa tulosvaihtoehtoa. Esimerkiksi tuotantoprosessissa syntyvä tuote luokitellaan yhteen ja vain yhteen seuraavista kategorioista: ensiluokkainen (1), sekunda (2) tai viallinen (3). Olkoot ensiluokkaisen, sekundan ja viallisen todennäköisyydet vastaavasti p_1, p_2 ja $p_3 = 1 - p_1 - p_2$. Valmistetaan n tuotetta. Olkoon X_1 = ensiluokkaisten lukumäärä, X_2 = sekundatuotteiden lukumäärä ja $X_3 = n - X_1 - X_2$ = viallisten lukumäärä tuote-erässä. Jos x_1 ja x_2 ovat sellaiset epänegatiiviset kokonaisluvut, että $x_1 + x_2 \le n$, niin todennäköisyys saada x_1 ensiluokkaista, x_2 sekundaa ja $n - x_1 - x_2$ viallista jossain annetussa järjestyksessä on

$$p_1^{x_1} p_2^{x_2} (1 - p_1 - p_2)^{n - x_1 - x2}.$$

Sellaisia n:n tuotteen järjestyksiä, joissa on x_1 ensiluokkaista, x_2 sekundaa ja $n - x_1 - x_2$ viallista, on

$$\binom{n}{x_1} \binom{n - x_1}{x_2} = \frac{n!}{x_1!\, x_2!\, (n - x_1 - x_2)!}$$

kappaletta. Siksi *trinomijakauman todennäköisyysfunktio* on

$$f(x_1, x_2) = \frac{n!}{x_1!\, x_2!\, (n - x_1 - x_2)!}\, p_1^{x_1} p_2^{x_2} (1 - p_1 - p_2)^{n - x_1 - x_2}, \tag{4.9.1}$$

missä $f(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$. Kun (X_1, X_2) noudattaa trinomijakaumaa parametrein n, p_1 ja p_2, merkitään

$$(X_1, X_2) \sim \mathrm{Tri}(n, p_1, p_2).$$

On helppo todeta, että $X_1 \sim \mathrm{Bin}(n, p_1)$ (X_1:n reunajakauma) ja $X_2 \sim \mathrm{Bin}(n, p_2)$.

Multinomijakauma voidaan johtaa samalla periaatteella kuin trinomijakauma. Toistetaan n kertaa koe, jossa on k toisensa poissulkevaa tulosvaihtoehtoa. Merkitään tulosvaihtoehtoja $1, 2, \ldots, k$ ja olkoon p_i = tuloksen i todennäköisyys ja X_i on tuloksen i lukumäärä n:n kokeen sarjassa. Silloin k-ulotteisen satunnaisvektorin $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ arvoalue on

$$S = \{\, (x_1, x_2, \ldots, x_k) \mid 0 \le x_i \le n,\ x_1 + x_2 + \cdots + x_k = n \,\}.$$

X_i:t ovat siis epänegatiivisia kokonaislukuarvoisia satunnaismuuttujia, joiden summa on n. Satunnaisvektori $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ noudattaa *k-ulotteista multinomijakaumaa* parametrein n ja $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)$, jota merkitään $\mathrm{Mult}(n, \boldsymbol{p})$. Multinomijakauman todennäköisyysfunktio on

$$f(x_1, x_2, \ldots, x_k) = \binom{n}{x_1\ x_2\ \ldots\ x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}, \tag{4.9.2}$$

missä $p_1 + p_2 + \cdots + p_k = 1$ ja $\binom{n}{x_1\ x_2\ \ldots\ x_k} = \frac{n!}{x_1!\, x_2! \cdots x_k!}$. Multinomilauseen 2.9 nojalla voidaan helposti osoittaa, että todennäköisyyksien (4.9.2) summa yli arvoalueen S on 1, joten kyseinen funktio on todellakin todennäköisyysfunktio. Multinomijakaumassa jokaisen X_i:n *reunajakauma on binomijakauma,* eli $X_i \sim \mathrm{Bin}(n, p_i)$, $i = 1, 2, \ldots, k$. Esitetään nyt multinomijakaumaa koskevat perustulokset lauseen muodossa.

Lause 4.14 *1. Funktio (4.9.2) on multinomijakauman todennäköisyysfunktio kaikilla positiivisilla kokonaisluvuilla n ja kaikilla sellaisilla p_1, $\ldots$, p_k, että $0 \le p_i \le 1$ ja $p_1 + \cdots + p_k = 1$.*

2. Jos $\boldsymbol{X} \sim \mathrm{Mult}(n, \boldsymbol{p})$, niin

$$X_i \sim \mathrm{Bin}(n, p_i) \qquad \textit{ja} \qquad (X_i, X_j) \sim \mathrm{Tri}(n, p_i, p_j),$$

3.

$$E(X_i) = np_i, \qquad \mathrm{Var}(X_i) = np_i(1 - p_i), \qquad \mathrm{Cov}(X_i, X_j) = -np_ip_j,$$
$$M(t) = E[\exp(t_1X_1 + \cdots + t_kX_k)] = \left(p_1 \mathrm{e}^{t_1} + \cdots + p_k \mathrm{e}^{t_k}\right)^n.$$

Multinomijakauma liittyy otantaan palauttaen. Olkoon uurnassa erivärisiä palloja yhteensä N, värien lukumäärä on k ja väriä i olevia palloja on N_i kappaletta ($i = 1, 2, \ldots, k$) ja $N_1 + N_2 + \cdots + N_k = N$. Otannassa palauttaen todennäköisyys p_i saada väri i on N_i/N jokaisessa nostossa. Valitaan uurnasta n palloa palauttaen ja olkoon X_i väriä i olevien pallojen lukumäärä otoksessa. Silloin satunnaismuuttujien X_1, X_2, $\ldots$, X_k yhteisjakauma on multinomijakauma.

Otannassa palauttamatta uurnan sisältö muuttuu ja siten myös valintatodennäköisyydet muuttuvat valintaprosessin aikana. Satunnaismuuttujien X_1, X_2, $\ldots$, X_k yhteisjakauman johtamiseksi meidän on yleistettävä alaluvussa 2.6.1 esitetty hypergeometrinen jakauma.

Olkoon x_i väriä i ($1 \le i \le k$) olevien pallojen lukumäärä otannassa palauttamatta. Millä todennäköisyydellä saadaan otos, jossa eri väriä olevien lukumäärät ovat $(x_1, x_2, \ldots, x_k)$? Koska uurnassa on N_i kappaletta väriä i, niin $0 \le x_i \le N_i$. Otoskoko on n ja $n = x_1 + x_2 + \cdots + x_k$. Nyt väriä 1 olevat x_1 palloa voidaan valita $\binom{N_1}{x_1}$ tavalla, väriä 2 olevat $\binom{N_2}{x_2}$ tavalla ja lopulta väriä k olevat pallot $\binom{N_k}{x_k}$ tavalla. Suotuisten otosten lukumäärä on tuloperiaatteen nojalla

$$\binom{N_1}{x_1}\binom{N_2}{x_2}\cdots\binom{N_k}{x_k}.$$

Koska kaikkien mahdollisten n:n kokoisten otosten lukumäärä on $\binom{N}{n}$, niin todennäköisyys saada lukumäärät $(x_1, x_2, \ldots, x_k)$ erivärisiä palloja on

$$f(x_1, x_2, \ldots, x_k) = \frac{\binom{N_1}{x_1}\binom{N_2}{x_2}\cdots\binom{N_k}{x_k}}{\binom{N}{n}}, \tag{4.9.3}$$

missä $x_1 + x_2 + \cdots + x_k = n$. Tämä on *moniulotteisen hypergeometrisen jakauman* todennäköisyysfunktio.

Diskreetit jakaumat: Yhteenveto

Bernoulli
Ber(p)

$$f(x) = p^x(1-p)^{1-x}, \qquad x = 0, 1$$
$$E(X) = p, \qquad \mathrm{Var}(X) = p(1-p)$$
$$M(t) = 1 - p + p\mathrm{e}^t$$

Binomi
Bin(n, p)

$$f(x) = \binom{n}{x} p^x(1-p)^{n-x}, \qquad x = 0, 1, \ldots, n$$
$$E(X) = np, \qquad \mathrm{Var}(X) = np(1-p)$$
$$M(t) = (1 - p + p\mathrm{e}^t)^n$$

Negatiivinen binomi
NBin(r, p)

$$f(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \qquad x = r, r+1, \ldots$$

$$E(X) = \frac{r}{p}, \qquad \mathrm{Var}(X) = \frac{r(1-p)}{p^2}$$

$$M(t) = \left[\frac{p\mathrm{e}^t}{1-(1-p)\mathrm{e}^t}\right]^r, \qquad t < -\log(1-p)$$

Geometrinen
Geo(p)

$$f(x) = p(1-p)^{x-1}, \qquad x = 1, 2, 3, \ldots$$

$$E(X) = \frac{1}{p}, \qquad \mathrm{Var}(X) = \frac{1-p}{p^2}$$

$$M(t) = \frac{p\mathrm{e}^t}{1-(1-p)\mathrm{e}^t}, \qquad t < -\log(1-p)$$

Hypergeometrinen
HGeo(n, N, p)

$$f(x) = \frac{\binom{Np}{x}\binom{N-Np}{n-x}}{\binom{N}{n}}, \qquad \begin{array}{l} X \le pN \quad \text{ja} \\ n - X \le N - Np \end{array}$$

$$E(X) = np, \qquad \mathrm{Var}(X) = \frac{N-n}{N-1} np(1-p),$$

Negatiivinen hypergeometrinen
NHGeo(r, N, p)

$$f(x) = \binom{x-1}{r-1} \frac{\binom{N-x}{Np-r}}{\binom{N}{Np}}, \qquad x = r, r+1, \ldots, N$$

$$E(X) = r \cdot \frac{N+1}{Np+1}$$

$$\mathrm{Var}(X) = \frac{r(1-p)N(N+1)(Np+1-r)}{(Np+1)^2(Np+2)}$$

Poisson
Poi(λ)

$$f(x) = \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, \ldots$$

$$E(X) = \lambda, \qquad \mathrm{Var}(X) = \lambda$$

$$M(t) = \exp(\lambda \mathrm{e}^t - \lambda), \qquad -\infty < t < \infty$$

Poissonin prosessi. Laskuriprosessi $\{\, N(t),\ t \ge 0 \,\}$, jonka intensiteetti λ.

1. $N(0) = 0$.
2. Prosessin lisäykset ovat riippumattomat.
3. Tapahtumien lukumäärä jokaisella t:n pituisella välillä noudattaa Poissonin jakaumaa, jonka odotusarvo on λt:

$$P[N(t+s) - N(s) = x] = \mathrm{e}^{-\lambda t}\frac{(\lambda t)^x}{x!}, \qquad x = 0, 1, \ldots$$

kaikilla $s, t \ge 0$.

Kaksiulotteiset jakaumat

Satunnaisvektorin (X, Y) yhteisjakauma.

Todennäköisyysfunktio $f(x, y)$ toteuttaa ehdot

1. $0 \leq f(x, y) \leq 1, \quad$ kaikilla $(x, y) \in S$ ja
2. $\sum\limits_{(x,y)\in S} f(x, y) = 1,$

missä S on satunnaisvektorin (X, Y) arvojoukko.

Reunajakaumien todennäköisyysfunktiot:

$$f_X(x) = \sum_{y\in S_Y} f(x, y), \quad x \in S_X; \qquad f_Y(y) = \sum_{x\in S_X} f(x, y), \quad y \in S_Y.$$

Ehdolliset todennäköisyysfunktiot:

$$f_1(x \mid y) = \frac{f(x, y)}{f_Y(y)} \qquad \text{ja} \qquad f_2(y \mid x) = \frac{f(x, y)}{f_X(x)}, \qquad (x, y) \in S.$$

Bernoulli
Ber2(p_1, p_2, p_{11})

$f(x, y) = p_{00}^{(1-x)(1-y)} p_{01}^{(1-x)y} p_{10}^{x(1-y)} p_{11}^{xy}, \qquad$ missä
$p_{00} + p_{01} + p_{10} + p_{11} = 1, \quad x \in \{0, 1\} \quad$ ja $\quad y \in \{0, 1\}$
$E(X) = p_1 = P(X = 1), \qquad E(Y) = p_2 = P(Y = 1)$
$E(XY) = p_{11} = P(X = 1, Y = 1)$

Multinomi
Mult$(n, \boldsymbol{p})$

$\boldsymbol{p} = (p_1, p_2, \ldots, p_k), \qquad$ missä
$0 \leq p_i \leq 1 \quad$ ja $\quad p_1 + p_2 + \cdots + p_k = 1$
$E(X_i) = np_i, \qquad \mathrm{Var}(X_i) = np_i(1 - p_i)$
$\mathrm{Cov}(X_i, X_j) = -np_ip_j$
$X_i \sim \mathrm{Bin}(n, p_i), \quad (X_i, X_j) \sim \mathrm{Mult}(n; p_i, p_j, 1 - p_i - p_j)$
$M(t) = (p_1\mathrm{e}^{t_1} + \cdots + p_k\mathrm{e}^{t_k})^n$

Hypergeometrinen

$f(x_1, x_2, \ldots, x_k) = \dfrac{\binom{N_1}{x_1}\binom{N_2}{x_2}\cdots\binom{N_k}{x_k}}{\binom{N}{n}}, \qquad$ missä
$N_1 + N_2 + \cdots + N_k = N \quad$ ja $\quad x_1 + x_2 + \cdots + x_k = n$

Harjoituksia

1. Olkoon X satunnaismuuttuja, jonka arvojoukko on $S_X = \{x_1, x_2\}$ ja todennäköisyysfunktio $P(X = x_1) = p$, $P(X = x_2) = 1 - p$.

(a) Laske $E(X^r)$, $r = 1, 2$ ja
(b) $\mathrm{Var}(X)$.

(c) Määritä X:n momenttifunktio.

2. Olkoon $X \sim \mathrm{Ber}(p)$ ja Y sellainen satunnaismuuttuja, jonka arvojoukko on $S_Y = \{y_1, y_2\}$ ja todennäköisyysfunktio $P(Y = y_1) = p$, $P(Y = y_2) = 1 - p$. Lausu Y satunnaismuuttujan X avulla.

3. Heitetään lanttia n kertaa (n riippumatonta Bernoullin koetta). Olkoon kruunun (R) todennäköisyys p ja X toistosten RR lukumäärä heittosarjassa.

(a) Mitä on $E(X)$? Mikä on $E(X)$:n arvo, kun kun $n = 200$?

(b) Laske $\mathrm{Var}(X)$.

(c) Mitä on toistosten RRRR lukumäärän odotusarvo?

(Vihje: Katso Esimerkki 4.2.)

4. Jos X noudattaa binomijakaumaa, jonka odotusarvo on 6 ja varianssi 2.4, niin mitä on $P(X = 5)$?

5. Hatussa on N yhdestä lähtien juoksevasti numeroitua arpalippua. Valitaan hatusta n:n arvan satunnaisotos palauttamatta (ks. Esimerkki 4.2). Olkoon X suurin valittujen arpalippujen järjestysnumeroista.

(a) Piirrä X:n todennäköisyys- ja kertymäfunktion kuvaajat, kun $N = 100$ ja $n = 10$.

(b) Piirrä X:n odotusarvon kuvaaja n:n funktiona, kun $N = 100$.

6. Valitaan satunnaisesti ja toisistaan riippumatta 2000 pistettä yksikköneliöstä $\{\,(x, y) \mid 0 \leq x \leq 1,\ 0 \leq y \leq 1\,\}$. Olkoon Z yksikköympyrään $\{\,(x, y) \mid x^2 + y^2 \leq 1\,\}$ osuvien pisteiden lukumäärä.

(a) Mitä jakaumaa Z noudattaa?

(b) Laske Z:n odotusarvo ja hajonta.

(c) Satunnaismuuttujan $\frac{Z}{500}$ odotusarvo?

(d) Generoi 2000 satunnaislukuparia. Määritä Z:n arvo ja laske sen avulla π:n likiarvo.

7. Erääseen 90:n virheettömän kännykän tuote-erään oli sekaantunut 10 viallista. Valitaan tästä 100:n kännykän joukosta 30 kännykän otos palauttamatta. Olkoon X viallisten lukumäärä otoksessa.

(a) Määritä X:n todennäköisyysfunktio.

(b) Laske $P(X = 10)$.

(c) Valitaan kännyköitä testaukseen satunnaisotannalla yksitellen palauttamatta, kunnes kaikki vialliset on löydetty. Olkoon Y tarvittavien testien lukumäärä. Laske $P(Y \geq 20)$, eli todennäköisyys, että tarvitaan ainakin 20 testiä.

8. Heitetään harhatonta lanttia, kunnes havaitaan toistos RR (kaksi kruunua peräkkäin). Olkoon X tarvittavien heittojen lukumäärä. Olkoon f_n n. Fibonaccin luku, joka määritellään siten, että $f_1 = f_2 = 1$ ja $f_n = f_{n-1} + f_{n-2}$, $n = 3, 4, \ldots$.

(a) Osoita, että X:n todennäköisyysfunktio on

$$f(x) = \frac{f_{x-1}}{2^x}, \qquad x = 2, 3, 4, \ldots$$

(b) Osoita tuloksen

$$f_x = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^x - \left(\frac{1-\sqrt{5}}{2}\right)^x\right]$$

avulla, että $\sum_{x=2}^{\infty} = 1$.

(c) Osoita, että $E(X) = 6$.

(d) Osoita, että $E[X(X-1)] = 52$ ja $\mathrm{Var}(X) = 22$.

(e) Simuloi X:n arvoja ja tarkastele, vastaavatko simuloinnin tulokset teoreettisia tuloksia.

9. Eräässä vaalissa 4000:sta äänestäjästä 100 kannatti ehdokasta A. Jos valitaan 50 alkion otos äänestäjistä esitutkimukseen, niin millä todennäköisyydellä haastatelluista korkeintaan 5 kannattaa A:ta?

10. Yritykseen tulee lähetys, joka sisältää 1000 varaosaa. Tarkistussuunnitelman mukaan $n = 100$ satunnaisesti valittua (palauttamatta) varaosaa on tarkistettava. Tuote-erä hyväksytään, jos tarkistuksessa ei löydy kahta viallista enempää. Mikä on todennäköisyys, että tuote-erä hyväksytään? Laske todennäköisyys

(a) hypergeometrisen jakauman avulla.

(b) Laske sitten sama todennäköisyys käyttäen hypergeometrisen jakauman likiarvona binomijakaumaa

(c) ja Poissonin jakaumaa.

11. Oletetaan, että $X \sim \mathrm{Poi}(\lambda)$. Osoita, että $E(X) = \lambda$ ja $\mathrm{Var}(X) = \lambda$.

12. Leipomossa valmistetaan suuri taikina, josta tehdään rusinaleivoksia. Leipoyrittäjä haluaa, että 95 % leivoksista sisältää ainakin 2 rusinaa. Kuinka monta rusinaa leivosta kohti hänen pitää sekoittaa taikinaan?

13. Laboratoriohiiriin ruiskutetaan kahta eri liuosta. Ensimmäisessä liuoksessa on keskimäärin c kappaletta C-tyypin organismeja millitrassa ja toisessa liuoksessa keskimäärin d kappaletta D-tyypin organismeja millitrassa. Organismit ovat jakautuneet nesteeseen täysin satunnaisesti. Jokaiseen hiireen ruiskutetaan kumpaakin liuosta yksi millitra. Hiiri säilyy hengissä jos ja vain jos kummassakaan ruiskeessa ei ole yhtään organismia.

(a) Millä todennäköisyydellä hiiri jää eloon?

(b) Millä todennäköisyydellä kuolleista hiiristä löytyy molempia organismeja?

(Vihje: Käytä Poissonin jakaumaa.)

14. Tehtaalla sattuu keskimäärin 1.5 onnettomuutta kuukaudessa. Määritä seuraavien tapahtumien todennäköisyydet:

(a) Ei onnettomuuksia tammikuussa,

(b) yhteensä neljä onnettomuutta helmikuussa ja maaliskuussa,

(c) ainakin yksi onnettomuus vuoden jokaisena kuukautena.

(Vihje: Käytä Poissonin jakaumaa.)

15. Olkoot X ja Y toisistaan riippumattomat Poissonin jakaumaa noudattavat satunnaismuuttujat. Olkoon $E(X) = 1$ ja $E(Y) = 2$.

(a) Laske todennäköisyys $P(X + Y) = 5$.

(b) Millä kokonaislukuarvolla n todennäköisyys $P(X+Y) = n$ saavuttaa maksiminsa?

(c) Lausu todennäköisyys $P(X + Y) = 5$ satunnaismuuttujien X ja Y todennäköisyysfunktioiden avulla.

16. Kirjassa on 200 sivua. Painovirheiden lukumäärä jokaisella sivulla noudattaa Poissonin jakaumaa, jonka keskiarvo on 0.01. Painovirheiden lukumäärät eri sivuilla ovat toisistaan riippumattomat.

(a) Mikä on virheettömien sivujen lukumäärän odotusarvo ja hajonta?

(b) Kirjan oikolukija havaitsee minkä tahansa annetun virheen todennäköisyydellä 0.9. Mikä on oikolukijan havaitsemien virheellisten sivujen lukumäärän odotusarvo?

17. Määritellään X:n ja Y: yhteisjakauman todennäköisyysfunktio seuraavasti:

$$f(x, y) = \frac{x + y}{32}, \qquad \text{kun} \quad x = 1, 2; \quad y = 1, 2, 3, 4.$$

Määritä

(a) X:n reunajakauman todennäköisyysfunktio ja

(b) Y:n reunajakauman todennäköisyysfunktio.

(c) Laske $P(X > Y)$,

(d) $P(Y = 2X)$ ja

(e) $P(X + Y = 3)$.

18. X:n ja Y:n yhteisjakauman todennäköisyysfunktio on määritelty 17. tehtävässä.

(a) Laske odotusarvot μ_X ja μ_Y,

(b) varianssit σ_X^2 ja σ_Y^2 sekä

(c) korrelaatiokerroin ρ.

(d) Ovatko X ja Y riippumattomat?

19. X:n ja Y:n yhteisjakauman todennäköisyysfunktio on määritelty 17. tehtävässä.

(a) Määritä X:n ehdolliset todennäköisyysfunktiot $f_1(x \mid y)$ ehdolla $y = 1, 2, 3$ ja $y = 4$.

(b) Määritä Y:n ehdolliset todennäköisyysfunktiot $f_2(y \mid x)$ ehdolla $x = 1$ ja $x = 2$.

(c) Laske $P(1 \leq Y \leq 3 \mid X = 1)$, $P(Y \leq 2 \mid X = 2)$, ja $P(X = 2 \mid Y = 3)$.

(d) Laske $E(Y \mid X = 1)$ ja $\mathrm{Var}(Y \mid X = 1)$.

20. Testataan kymmenen suojakypärän iskukestävyys. Kypärät jaetaan kahteen viiden ryhmään. Ensimmäisen ryhmän kypärille annetaan isku, joka särkee kypärän todennäköisyydellä 0.1. Toisen ryhmän kypäriä isketään voimalla, joka särkee kypärän todennäköisyydellä 0.3. Millä todennäköisyydellä ensimmäisen ryhmän kypäriä rikkoontuu enemmän kuin toisen ryhmän kypäriä?

21. Oletetaan, että satunnaismuuttujat X_1, X_2, X_3 noudattavat multinomijakaumaa $\mathrm{Mult}(5; 0.1, 0.3, 0.6)$ [Toisin sanoen $(X_1, X_2) \sim \mathrm{Tri}(5; 0.1, 0.3)$].

(a) Määritä X_1:n reunajakauma ja X_2:n

(b) ehdollinen todennäköisyysfunktio $f_2(x_2 \mid x_1 = 1)$.

22. Nostetaan tavallisesta korttipakasta (52 korttia) satunnaisesti palauttamatta 13 korttia. Olkoon X_1 patojen lukumäärä, X_2 herttojen lukumäärä ja $13 - X_1 - X_2$ ruutujen ja ristien lukumäärä otoksessa.

(a) Määritä X_1:n ja X_2:n yhteisjakuman todennäköisyysfunktio. Mikä on satunnaisvektorin (X_1, X_2) arvoalue?

(b) X_1:n todennäköisyysfunktio ja arvoalue?

23. Oletetaan, että (X, Y) noudattaa trinomijakaumaa $\mathrm{Tri}\big(3, \frac{1}{6}, \frac{1}{2}\big)$.

(a) Laske odotusarvot μ_X ja μ_Y,

(b) varianssit σ_X^2 ja σ_Y^2 sekä

(c) kovarianssi $\mathrm{Cov}(X, Y)$ ja

(d) korrelaatiokerroin ρ.

24. Satunnaismuuttujat $X \geq 0$ ja $Y \geq 0$ ovat riippumattomat ja saavat vain kokonaislukuarvoja. Osoita, että

(a) $P(X+Y=n) = \sum_{k=0}^{n} P(X=0)\, P(Y=n-k)$.

(b) Heitetään 4:ää harhatonta noppaa ja lasketaan silmälukujen summa. Mikä on todennäköisyys, että summa on 8? (Vihje: Olkoon X kahden nopan silmälukujen summa ja Y kahden muun nopan silmälukujen summa.)