Todennäköisyyslaskentaa

Todennäköisyyden käsite

1) Heitetään arpanoppaa. Mikä on todennäköisyys, että saadaan ykkönen?
2) Mikä on todennäköisyys, että syntyvä lapsi on tyttö?
3) Sääennusteesssa sanotaan, että huomenna sataa 50 %:n todennäköisyydellä. Mitä tämä tarkoittaa?
4) Todennäköisesti menen illalla elokuviin. Mitä tämä tarkoittaa?

1) Jos arpanoppa on symmetrinen, jokaisen silmäluvun todennäköisyys voidaan olettaa samaksi. N heiton sarjassa ykkösten lukumäärä $n(1)$ on likimain $N/6$. Tarkemmin sanottuna

$$\lim_{N\to\infty} \frac{n(1)}{N} = \frac{1}{6}.$$

Tämän perusteella todennäköisyys saada ykkönen on

$$P(1) = \frac{1}{6}.$$

Tämä on todennäköisyyden *frekvenssitulkinta*.

Arpanopan, korttipakan tms. tapauksessa todennäköisyys voidaan päätellä symmetriaominaisuuksien avulla. Todennäköisyyksiä voidaan laskea kombinatoriikan ja erilaisten jakaumien avulla.

2) Tyttöjen ja poikien syntymätodennäköisyyksien ei tarvitse olla samoja, joten niitä ei voi päätellä symmetrian avulla. Todennäköisyyttä voidaan arvioida tilastollisesti.

Jos kaikkiaan N syntyneestä lapsesta $n(\mathrm{t})$ on tyttöjä, todennäköisyys saada tyttö on $n(\mathrm{t})/N$. Kun N kasvaa, tämä lähenee todennäköisyyttä $P(\mathrm{t})$.

3) Huomisia on vain yksi, joten frekvenssejä ei voi laskea. Voidaan kuitenkin laskea suuri määrä ennusteita hieman häirityillä alkuarvoilla ja katsoa, kuinka monessa niistä sataa.

4) Kyseessä on *subjektiivinen todennäköisyys*, jota ei yleensä voi ennustaa.

Matemaattinen teoria

Ongelmia aiheutuu usein siitä, ettei ole määritelty täsmällisesti, millaisen tapahtuman todennäköisyyttä lasketaan.

Todennäköisyyslaskennan täsmällisen aksiomaattisen teorian esitti Andrei Kolmogorov 1929.

Kokeen jokainen mahdollinen tulos muodostaa yhden *alkeistapauksen*. Esimerkiksi nopanheiton yksi alkeistapaus on 'saadaan ykkönen'.

Kaikkien alkeistapausten joukko on *perusjoukko* Ω. Nopanheiton perusjoukko on $\{1, 2, 3, 4, 5, 6\}$.

Perusjoukon osajoukot ovat *tapahtumia*. Noppaa heitettäessä mahdollisia tapahtumia ovat esimerkiksi 'saadaan ykkönen', 'saadaan parillinen silmäluku'. Jos perusjoukko on äärellinen, tapahtumien joukko $\mathcal{F}$ on usein sama kuin perusjoukon kaikkien osajoukkojen joukko.

Yleisessä tapauksessa vaaditaan, että $\mathcal{F}$ on σ-algebra:

(1) $\Omega \in \mathcal{F}$.

(2) Jos $A \in \mathcal{F}$, myös $\Omega - A \in \mathcal{F}$.

(3) Jos $A_i \in \mathcal{F}$, $i = 1, 2, ...$, $\cup_i A_i \in \mathcal{F}$.

Näiden lisäksi tarvitaan kuvaus $P : \mathcal{F} \to \mathbf{R}$. P on todennäksöisyys, jos se toteuttaa aksioomat

(1) $P(A) \geq 0$ kaikille $A \in \mathcal{F}$.

(2) $P(\Omega) = 1$.

(3) Jos $A_i \in \mathcal{F}$, $i = 1, 2, ...$ ja $A_i \cap A_j = \emptyset$ kaikilla $i \neq j$, $P(\cup_i A_i) = \sum_i P(A_i)$.

Todennäköisyysavaruus on nyt kolmikko $(\Omega, \mathcal{F}, \mathcal{P})$.

Teoria ei puutu siihen, miten alkeistapausten todennäköisyydet lasketaan.

Aksioomista lähtien voidaan osoittaa mm. seuraavat todennäköisyyden ominaisuudet:

$$
\begin{aligned}
&P(\emptyset) = 0,\\
&0 \leq P(A) \leq 1,\\
&P(\Omega - A) = 1 - P(A),\\
&P(A - B) = P(A) - P(A \cap B),\\
&P(A \cup B) = P(A) + P(B) - P(A \cap B),\\
&A \subset B \Rightarrow P(A) \leq P(B).
\end{aligned}
$$

Kombinatoriikkaa

Kun perusjoukko on äärellinen ja alkeistapaukset symmetrisiä, todennäköisyydet voidaan laskea luettelemalla kaikki vaihtoehdot. Eri tapausten lukumääriä voidaan laskea kombinatoriikan avulla.

Permutaatiot

Kuinka monella eri tavalla kolme kirjainta abc voidaan järjestää? Ensimmäinen kirjain voi olla mikä tahansa kolmesta, toinen jompikumpi jäljellejääneistä ja kolmas ainoa jäljellejäänyt. Vaihtoehtoja on $3 \times 2 \times 1 = 6$ kappaletta.

Yleisemmin n alkiota voidaan järjestää

$$1 \times 2 \times 3 \times \cdots \times n = n!$$

tavalla. Tämä on n:n alkion permutaatioiden määrä.

Kun n on suuri, n-kertomalle voidaan käyttää Stirlingin approksimaatiota. Olkoon

$$f(n) = \left(\frac{n}{\mathrm{e}}\right)^n \sqrt{2\pi n}.$$

Silloin

$$\lim_{n\to\infty} \frac{f(n)}{n!} = 1.$$

Huomaa, että suhde lähestyy asymptoottisesti ykköstä, mutta erotus $|n! - f(n)|$ kasvaa rajatta, joten myös approksimaation absoluuttinen virhe kasvaa.

Tuloperiaate

Jos kaksivaiheisen kokeen ensimmäinen osa voidaan suorittaa n tavalla ja toinen m tavalla, erilaisia tapoja suorittaa koko koe on nm kappaletta.

Yleisemmin: jos koe voidaan suorittaa k vaiheessa ja vaiheessa $i = 1, \ldots, k$ erilaisia vaihtoehtoja on n_i kappaletta, koko kokeen erilaisten tulosvaihtoehtojen määrä on

$$n_1 n_2 \cdots n_k.$$

Esimerkki: Veikkausrivissä on 13 kohdetta, joista kussakin on kolme vaihtoehtoa (1, x, 2). Erilaisia rivejä on siten

$$3 \times 3 \times \cdots 3 = 3^{13} = 1594323$$

Summaperiaate

Jos koe voidaan suorittaa k tavalla, jotka ovat toisensa poissulkevia, ja tavalla i on n_i mahdollisuutta, kaikkien tulosvaihtoehtojen määrä on

$$n_1 + n_2 + \cdots n_k.$$

Esimerkki: Kuinka monta erilaista kahden tai kolmen merkin merkkijonoa voidaan muodostaa 28 aakkosen avulla?

Kahden merkin jonoja on $28^2 = 784$ kappaletta ja kolmen merkin jonoja $28^3 = 21952$ kappaletta, joten erilaisia merkkijonoja on $784 + 21952 = 22736$.

Variaatiot

Jos meillä on N erilaista oliota, niistä voidaan poimia erilaisia k:n olion järjestettyjä jonoja kaikkiaan

$$(N)_k = N(N-1)(N-2)\cdots(N-k+1) = \frac{N!}{(N-k)!}$$

tavalla. Tämä suure, *k-variaatio*, on N:stä kappaleesta k:ttain otettujen variaatioiden määränä.

Kombinaatiot

Jos sisäisestä järjestyksestä ei välitetä, N:n olion joukosta voidaan poimia erilaisia k:n alkion joukkoja

$$\binom{N}{K} = \frac{(N)_k}{k!} = \frac{N!}{k!(N-k)!}$$

kappaletta. $\binom{N}{k}$ on binomikerroin.

Esimerkiksi tavallisesta korttipakasta voidaan jakaa erilaisia pokerikäsiä

$$\binom{52}{5} = \frac{52!}{5!47!} = \frac{48 \times 49 \times 50 \times 51 \times 52}{1 \times 2 \times 3 \times 4 \times 5} = 2598960$$

kappaletta.

Otanta ilman takaisinpanoa

Oletetaan, että laatikossa on N palloa, joista V on valkeita ja $N - V$ mustia. Poimitaan laatikosta n palloa. Millä todennäköisyydellä saadaan v valkeaa palloa?

Kaikkiaan alkeistapauksia on $\binom{N}{n}$ kappaletta.

Valkeat pallot voidaan poimia $\binom{V}{v}$ tavalla ja mustat $\binom{N-V}{n-v}$ tavalla. Suotuisten alkeistapausten osuus on siten

$$P(A_v) = \frac{\binom{V}{v}\binom{N-V}{n-v}}{\binom{N}{n}}.$$

Otanta takaisinpanolla

Kun pallon väri on katsottu, se palautetaan laatikkoon.

Otos on n alkion jono, joita on kaikkiaan N^n kappaletta.

Valkeiden pallojen paikat voidaan valita jonosta $\binom{n}{v}$ tavalla.

Näille paikoille voidaan valita valkeat pallot V^v tavalla.

Muille paikoille voidaan valita mustat pallot $(N - V)^{n-v}$ tavalla.

$$P(A_v) = \binom{n}{v} \frac{V^k (N - V)^{n-v}}{N^n}.$$

Kun otos on pieni joukon kokoon N verrattuna, nämä todennäköisyydet ovat likimain samoja.

Ehdollinen todennäköisyys

Tiedämme, että on tapahtunut jokin tapahtuma B, $P(B) > 0$. Tapahtuman A todennäköisyys ehdolla B on

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Tästä saadaan kertolaskukaava

$$P(A \cap B) = P(B)P(A|B).$$

Esimerkki: Tiedetään, että noppaa heitettäessä saatiin korkeintaan kolmonen. Mikä on todennäköisyys, että silmäluku on pariton?

$A = \{1, 3, 5\}$, $B = \{1, 2, 3\}$.

$P(B) = P(\{1, 2, 3\}) = 1/2$.

$P(A \cap B) = P(\{1, 3\}) = 1/3$.

$P(A|B) = \frac{1/3}{1/2} = \frac{2}{3}$.

Riippumattomuus

Tapahtumat A ja B ovat riippumattomia ($A \perp\!\!\!\perp B$), jos $P(A \cap B) = P(A)P(B)$.

Ehdollisen todennäköisyyden määritelmän perusteella tämä on yhtäpitävää sen kanssa, että $P(A|B) = P(A)$, jos $P(B) > 0$.

Tieto tapahtumasta B ei vaikuta tapahtuman A todennäköisyyteen.

Em. esimerkissä

$$P(A \cap B) = 1/3,$$
$$P(A)P(B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4},$$

joten tapahtumat eivät ole riippumattomia.

Riippuvuus EI tarkoita, että tapahtumilla olisi kausaalinen yhteys.

Toistokokeet

Toistetaan koetta, joka voi päättyä kahdella tavalla:

$P(\text{"koe onnistuu"}) = P(A) = p$

$P(\text{"koe ei onnistu"}) = P(\Omega - A) = 1 - p = q$

Jos kokeet ovat riippumattomia, n kokeen sarjasta k kappaletta onnistuu todennäköisyydellä $p^k q^{n-k}$.

k onnistunutta koetta voidaan valita n kokeen sarjasta $\binom{n}{k}$ tavalla. Todennäköisyys, että koe onnistuu tasan k kertaa on siten

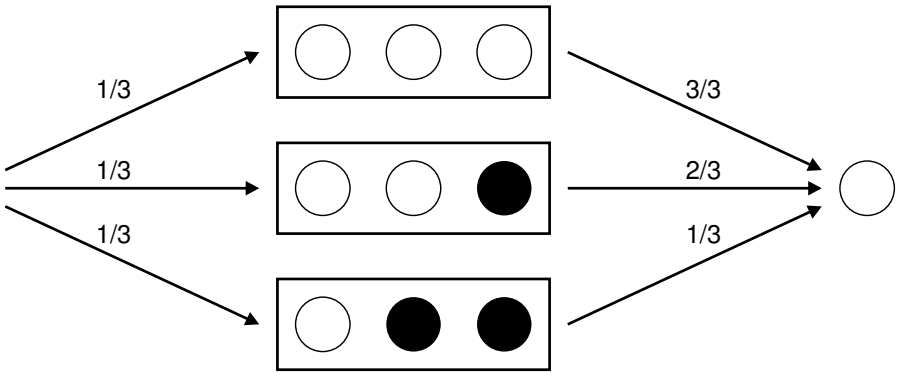
$$\binom{n}{k} p^k q^{n-k}.$$

Esimerkki: Heitetään noppaa 10 kertaa. Todennäköisyys, että saadaan kaksi kuutosta on

$$\binom{10}{2} (1/6)^2 (5/6)^8 \approx 0.29.$$

Kokonaistodennäköisyys

Valitaan satunnaisesti jokin allaolevista laatikoista ja nostetaan laatikosta satunnainen pallo. Millä todennäköisyydellä saadaan valkea pallo?



Merkitään A_i ='valitaan laatikko i' ja B ='saadaan valkea pallo'.

$P(A_1 \cap B) = \frac{1}{3}\frac{3}{3} = \frac{3}{9}$,

$P(A_2 \cap B) = \frac{1}{3}\frac{2}{3} = \frac{2}{9}$,

$P(A_3 \cap B) = \frac{1}{3}\frac{1}{3} = \frac{1}{9}$.

$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B) = \frac{6}{9} = \frac{2}{3}$.

Oletetaan, että perusjoukko ositetaan erillisiin tapahtumiin A_i, $A_i \cap A_j = \emptyset$, kun $i \neq j$, ja $\cup_i A_i = \Omega$. Silloin jokainen tapahtuma B voidaan jakaa osiin $B = \cup_i (A_i \cap B)$.

Tapahtuman B kokonaistodennäköisyys on nyt

$$P(B) = \sum_i P(A_i)P(B|A_i).$$

Bayesin kaava

Ehdollinen todennäköisyys oli

$$P(A_k|B) = \frac{P(A_k \cap B)}{P(B)}.$$

Osoittaja saadaan kertolaskukaavasta ja nimittäjään sijoitetaan kokonaistodennäköisyyden kaava. Saadaan Bayesin kaava

$$P(A_k|B) = \frac{P(A_k)P(B|A_k)}{\sum_i P(A_i)P(B|A_i)}.$$

Oletetaan, että edellisessä esimerkissä nostettiin valkea pallo. Mikä on todennäköisyys, että se on peräisin laatikosta 1?

Nyt $P(A_1) = 1/3$ ja $P(B|A_1) = 1$, joten

$$P(A_1|B) = \frac{1/3 \times 1}{2/3} = \frac{1}{2}.$$

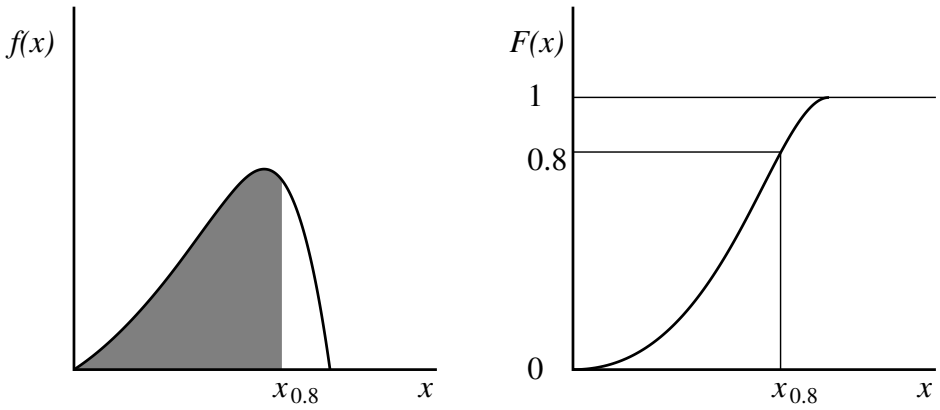
"Käänteinen päättely", yhteys tilastolliseen inversioteoriaan. Havaitusta tapahtumasta (tai sen todennäköisyysjakaumasta) päätellään sen aiheuttajan todennäköisyys.

Luvut $P(A_k)$ ovat *prioritodennäköisyyksiä* (jotakin ennalta tunnettua) ja luvut $P(A_k|B)$ *posterioritodennäköisyyksiä*. On siis havaittu B ja päätellään, millä todennäköisyydellä A_k on sen aiheuttaja.

Todennäköisyysjakaumat

Satunnaismuuttuja on suure, joka voi saada erilaisia arvoja satunnaisesti tietyillä todennäköisyyksillä.

Satunnaismuuttujaa kuvaa sen todennäköisyysjakauma.



Kertymäfunktio $F(x)$ ilmoittaa todennäköisyyden, että satunnaismuuttujan X arvo on korkeintaan x:

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

Kertymäfunktio on kasvava funktio (ei välttämättä aidosti kasvava).

Diskreetin jakauman *tiheysfunktio* $f(x)$ ilmoittaa, millä todennäköisyydellä muuttuja saa täsmälleen arvon x (pistetodennäköisyys):

$$f(x) = P(X = x).$$

Jatkuvan muuttujan tapauksessa tiheysfunktio voidaan kuitenkin määritellä siten, että

$$F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y.$$

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1.$$

Jos F on derivoituva, on

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}.$$

Todennäköisyys, että satunnaismuuttujan arvo on jollakin välillä, saadaan integroimalla tiheysfunktiota

$$P(x_0 \leq X \leq x_1) = \int_{x_0}^{x_1} f(x)\,\mathrm{d}x.$$

Jakaumien tunnusluvut

Odotusarvo $E(X)$ ilmoittaa muuttujan keskimääräisen arvon. Diskreetille jakaumalle odotusarvo on

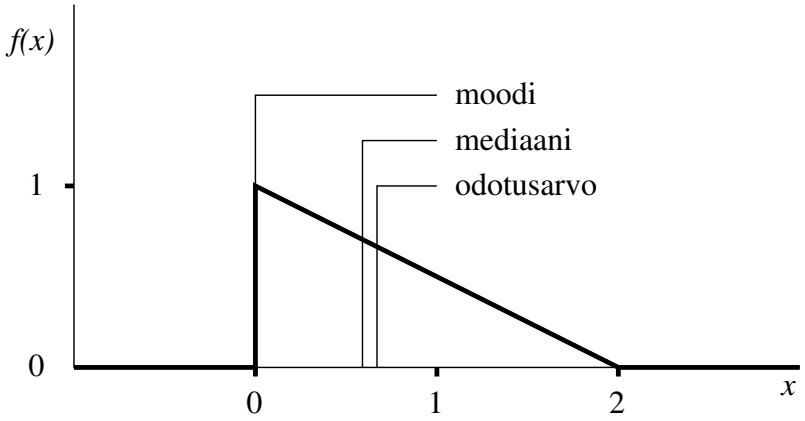
$$E(X) = \sum xP(X = x) = \sum xf(x),$$

ja jatkuvalle jakaumalle

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x.$$

Odotusarvo on lineaarinen operaattori:

$$E(aX + bY) = aE(X) + bE(Y).$$



Moodi ilmoittaa jakauman maksimikohdan eli satunnaismuuttujan kaikkein todennäköisimmän arvon.

Mediaani on raja, jota pienemmät ja suuremmat arvot esiintyvät molemmat samalla todennäköisyydellä.

Yleisessä tapauksessa ootusarvo, mediaani ja moodi voivat poiketa toisistaan.

Symmetrisille jakaumille odotusarvo ja mediaani ovat samoja. Jos symmetrisellä jakaumalla on vain yksi maksimi, myös moodi on sama kuin odotusarvo.

Jakauman *p-fraktiili* x_p on piste, jossa $F(x_p) = p$.

Varianssi kuvaa jakauman leveyttä:

$$\sigma^2 X = E(X - \bar{X})^2 = \int_{-\infty}^{\infty} (x - \bar{X})^2 f(x)\,\mathrm{d}x,$$

missä $\bar{X}$:llä on merkitty satunnaismuuttujan X odotusarvoa.

Hajonta eli *standardipoikkeama* on $\sigma = \sqrt{\sigma^2}$. Hajonta sopii hyvin kuvaamaan esimerkiksi mittaustulosten virheitä, koska sillä on sama dimensio kuin itse satunaismuuttujalla.

Varianssi ei ole lineaarinen operaattori. Sen sijaan sille pätee

$$\sigma^2(aX + b) = a^2\sigma^2 X.$$

Varianssin ominaisuuksien avulla voidaan johtaa mielivaltaisen funktion varianssia koskeva likimääräinen *virheiden kasaantumislaki* (law of the propagation of errors).

$$\sigma^2 g(X) \approx \left(\frac{\partial g}{\partial X}\right)_{X=EX} \sigma^2 X.$$

Varianssi on yksi muuttujan *momenteista* keskiarvonsa suhteen. Yleisesti n:s momentti on

$$\mu_n = E(X - \bar{X})^n.$$

Usein merkitään $\mu = E(X)$.

Diskreettejä jakaumia

Binomijakauma

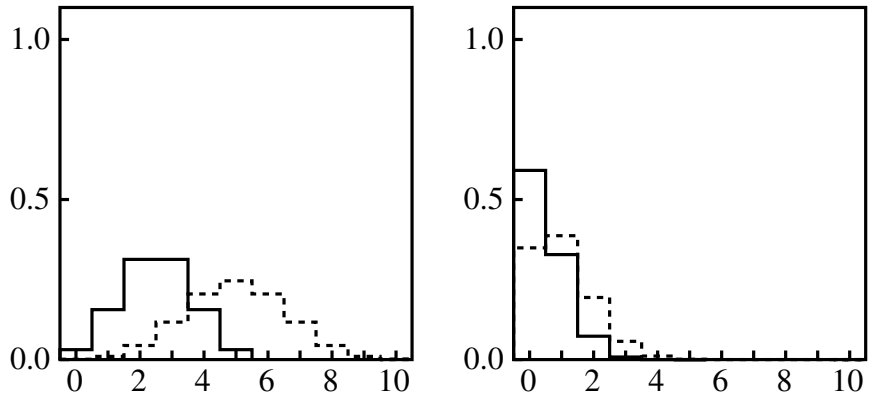
Oletetaan, että yksittäisessä kokeessa jonkin tuloksen todennäköisyys on p. Jos koe toistetaan N kertaa, tämä tulos saadaan tasan k kertaa todennäköisyydellä

$$P(k) = \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k} = \binom{N}{k} p^k (1-p)^{N-k}.$$

Binomijakaumaa noudattavan muuttujan odotusarvo ja varianssi ovat

$$\mu = Np,$$
$$\sigma^2 = Npq.$$

Kun toistojen määrä N kasvaa todennäköisyyden p pysyessä kiinteänä, binomijakauma lähenee normaalijakaumaa.



Binomijakauman tiheysfunktioita. Vasemmalla $p = 0.5$ (symmetrinen jakauma); oikealla $p = 0.1$. Yhtenäinen viiva $N = 5$, katkoviiva $N = 10$.

Poissonin jakauma

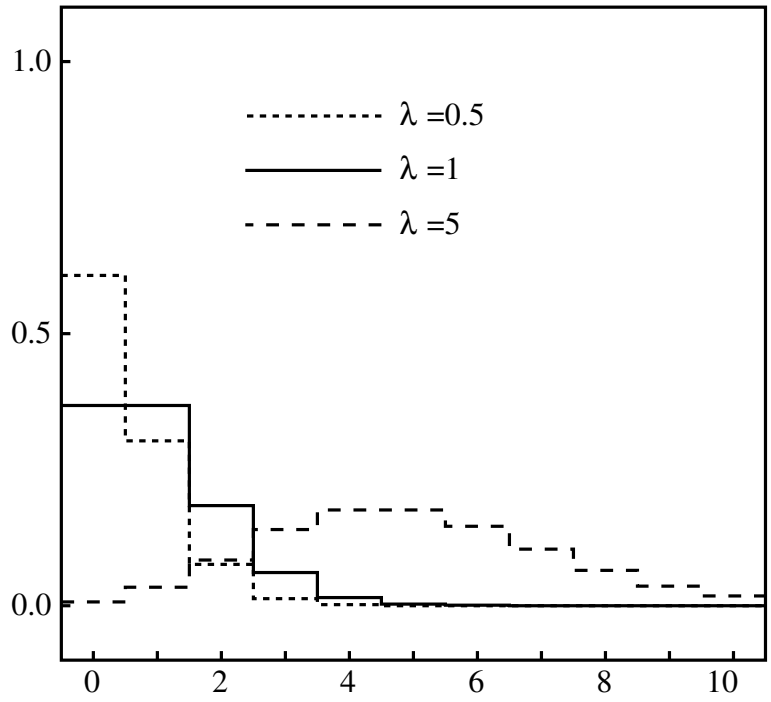
Merkitään $p = \lambda/N$. Jos toistojen kasvaessa p samalla pienenee siten, että Np pysyy vakiona, jakauma lähenee Poissonin jakaumaa

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

Odotusarvo ja varianssi ovat

$$\mu = \sigma^2 = \lambda.$$

Myös kaikki korkeammat momentit odotusarvon suhteen ovat λ:n suuruisia.



1) Jos tähdet ovat jakautuneet tasaisesti tietylle taivaan alueelle ja alue jaetaan samankokoisiin ruutuihin, yhteen ruutuun osuvien tähtien määrä noudattaa Poissonin jakaumaa.

2) Jos tähdestä tulevat fotonit jakautuvat tasaisesti ajan suhteen, aikayksikössä saapuvien fotonien määrä noudattaa Poissonin jakaumaa.

Esimerkki: CCD-kuvassa on 100 tähteä. Kuva jaetaan 25 samankokoiseen ruutuun. Millä todennäköisyydellä ruudussa on korkeintaan 2 tähteä?

Nyt odotusarvo on $\lambda = \mu = 100/25 = 4$.

$$P(0) = \frac{4^0}{0!}\mathrm{e}^{-4} \approx 0.018,$$
$$P(1) = \frac{4^1}{1!}\mathrm{e}^{-4} \approx 0.073,$$
$$P(2) = \frac{4^2}{2!}\mathrm{e}^{-4} \approx 0.147,$$

joten $P(x \leq 2) = P(0) + P(1) + P(2) = 0.238$.

Jatkuvia jakaumia

Tasainen jakauma

Muuttuja X on jakautunut tasaisesti välille (a, b), jos sen tiheysfunktio on

$$f(x) = \begin{cases} 1/(b-a), & \text{jos } a < x < b, \\ 0 & \text{muuten} \end{cases}$$

Tietokoneohjelmien satunnaislukugeneraattorit tuottavat yleensä lukuja, jotka ovat jakautuneet tasaisesti. Fortranissa tulos on välillä [0,1) oleva reaaliluku, C:ssä kokonaisluku välillä [0, RAND_MAX].

Mielivaltaista jakaumaa noudattavia satunnaislukuja z saadaan ratkaisemalla z yhtälöstä

$$F(z) = x,$$

missä x on tasaisen jakauman satunnaisluku ja F halutun jakauman kertymäfunktio.

Monille usein esiintyville jakaumille on kehitetty myös paljon tehokkaampia menetelmiä.

Eksponenttijakauma

Eksponenttijakauma kuvaa ilmiötä, joka ”ei muista menneisyyttään”.

Esimerkiksi säteilyn kulku väliaineessa. Fotoni absorboituu tietyllä tn:llä ensimmäisen metrin matkalla. Jos se selviää perille, sillä on sama tn selviytyä seuraavasta metristä. Todennäköisyys selviytyä matka s on verrannollinen lausekkeeseen $\mathrm{e}^{-s/\lambda}$, missä λ on keskimääräinen vapaa matka.

Eksponenttijakauman tiheysfunktio on

$$f(x) = ke^{-kx}$$

ja kertymäfunktio

$$F(x) = 1 - e^{-kx}.$$

Eksponenttijakauman odotusarvo ja varianssi ovat

$$\mu = \frac{1}{k}.$$
$$\sigma^2 = \frac{1}{k^2}.$$

Poissonin prosessissa tietyllä aikavälillä t tapahtuvien tapahtumien lukumäärä N noudattaa Poissonin jakaumaa siten, että odotusarvo on verrannollinen aikavälin pituuteen: $\mu = \lambda t$. Kahden tapahtuman väliaika T on satunnaismuuttuja, jonka kertymäfunktio on

$$F(t) = P(T \leq t).$$

Todennäköisyys, että ensimmäinen tapahtuma tapahtuu viimeistään hetkellä t on toisaalta sama kuin todennäköisyys, että kyseisellä aikavälillä tapahtuu ainakin yksi tapahtuma. Koska tämä tapahtumien määrä noudattaa Poissonin jakaumaa, saamme

$$F(t) = P(N \geq 1) = 1 - P(N = 0) = 1 - e^{-\lambda t},$$

josta derivoimalla saamme tiheysfunktioksi

$$f(t) = \lambda e^{-\lambda t}.$$

Tapahtumien välit noudattavat siis eksponenttijakaumaa.

Normaalijakauma

Normaalijakauman tiheysfunktio on

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

missä μ on odotusarvo ja σ^2 varianssi. Erityisesti puhutaan $(0,1)$-normaalijakaumasta, jos $\mu = 0$ ja $\sigma^2 = 1$. Sen tiheysfunktio on

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

Mikä tahansa normaalijakautunut muuttuja X voidaan aina palauttaa $(0,1)$-normaalijakaumaan muunnoksella

$$Z = \frac{X - \mu}{\sigma},$$

joten voimme keskittyä pelkästään $(0,1)$-normaalijakaumaan tulosten yleisyyden siitä kärsimättä.

Keskeinen raja-arvolause: Jos X_i:t, $i = 1, \ldots, n$ ovat riippumattomia satunnaismuuttujia, joista jokaisen odotusarvo on a ja varianssit σ_i^2, niiden keskiarvo

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

lähenee $(a, \sum \sigma_i^2/n)$-normaalijakaumaa, kun $n \to \infty$.

Normaalijakauman kertymäfunktio

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \, \mathrm{d}t$$

ei ole laskettavissa analyyttisesti. Useista aliohjelmakirjastoista löytyy rutiini error-funktion erf laskemiseen

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} \, \mathrm{d}t.$$

Normaalijakauman kertymäfunktio saadaan tämän avulla:

$$F(x) = \frac{1}{2} \left(1 + \operatorname{erf} \left(\frac{x}{\sqrt{2}} \right) \right).$$

Kertymäfunktiolle on myös useita approksimaatiokaavoja (ks. esim. Abramowitz ja Stegun).

$$t = \frac{1}{1 + 0.2316419x},$$

$$F(x) = 1 - \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \big(0.319381530t - 0.356563782t^2 + 1.781477937t^3 - 1.821255978t^4 + 1.330274429t^5 \big)$$

Tämän virhe on korkeintaan 7.5×10^{-8}.

Esimerkki: Mikä on todennäköisyys, että normaalijakautunut satunnaismuuttuja saa arvon, joka on suurempi kuin $\mu + 3\sigma$?

Normitettu (0,1)-normaalijakautunut muuttuja on $Z > 3$. Todennäköisyys on

$$1 - F(3) = 1 - \frac{1}{2} \left(1 + \operatorname{erf} \left(\frac{3}{\sqrt{2}} \right) \right) = 1 - 0.9986 = 0.0014.$$

Todennäköisyys, että satunnaismuuttujan arvo olisi yli 3σ:n verran suurempi kuin odotusarvo on siten runsaat 0.1 %.

χ^2-jakauma

Jos muuttujat X_i, $i = 1, \ldots, k$ ovat $(0, 1)$-normaalijakautuneita, niiden neliöiden summa

$$\chi_k^2 = \sum_{i=1}^{k} X_i^2$$

noudattaa χ^2-jakaumaa, jolla on k vapausastetta. Jakauman tiheysfunktio on

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2},$$

missä Γ on gammafunktio.

Studentin t-jakauma

Jos X on $(0, 1)$-normaalijakautunut ja Y noudattaa χ^2-jakaumaa k:lla vapausasteella, satunnaismuuttuja

$$t = \frac{X}{\sqrt{Y/k}}$$

noudattaa Studentin t-jakaumaa k:lla vapausasteella. Studentin jakauman tiheysfunktio on

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma(k/2)} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}.$$

Tässä ensimmäinen tekijä on pelkkä normitusvakio.

F-jakauma

Jos X noudattaa χ^2-jakaumaa m:llä vapausasteella ja Y noudattaa χ^2-jakaumaa k:lla vapausasteella, suhde

$$F_{mk} = \frac{X/m}{Y/k}$$

noudattaa F-jakaumaa vapausasteilla m ja k. Tälle jakaumalle saadaan tiheysfunktio

$$f(x) = \frac{\Gamma\left(\frac{k+m}{2}\right)\left(\frac{m}{k}\right)^{m/2}}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} x^{\frac{m}{2}-1}\left(1+\frac{mx}{k}\right)^{\frac{k+m}{2}}.$$

Useamman muuttujan jakaumat

Analogisesti yhden muuttujan jakauman kanssa voimme määritellä kertymäfunktion

$$F(x, y) = P(X \leq x \text{ ja } Y \leq y).$$

Jos kertymäfunktio on derivoituva, saamme tiheysfunktion

$$f(x, y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}F(x, y).$$

Tiheysfunktion avulla lausuttuna kertymäfunktio on

$$F(x, y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(x', y')\,\mathrm{d}x'\,\mathrm{d}y'.$$

Integroimalla toisen muuttujan yli saadaan reuna- eli *marginaalijakaumien* kertymäfunktiot

$$F(x) = \int_{-\infty}^{x}\left(\int_{-\infty}^{\infty} f(x', y')\,\mathrm{d}y'\right)\,\mathrm{d}x'$$

$$F(y) = \int_{-\infty}^{y}\left(\int_{-\infty}^{\infty} f(x', y')\,\mathrm{d}x'\right)\,\mathrm{d}y'.$$

Odotusarvo:

$$EX = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x' f(x', y')\,\mathrm{d}x'\,\mathrm{d}y',$$

$$EY = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y' f(x', y')\,\mathrm{d}x'\,\mathrm{d}y'.$$

Varianssit:

$$\sigma^2 X = E(X - \bar{X})^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x' - \bar{X})^2 f(x', y')\,\mathrm{d}x'\,\mathrm{d}y'.$$

$$\sigma^2 Y = E(Y - \bar{Y})^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y' - \bar{Y})^2 f(x', y')\,\mathrm{d}x'\,\mathrm{d}y'.$$

Korrelaatio, kovarianssi

Kaksiulotteisen jakauman uusi tunnusluku on kovarianssi:

$$\operatorname{cov}(X, Y) = E[(X - EX)(Y - EY)].$$

Kovarianssi voidaan laskea tiheysfunktion avulla:

$$\operatorname{cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x' - \bar{X})(y' - \bar{Y}) f(x', y') \, \mathrm{d}x' \, \mathrm{d}y'.$$

Jos muuttujat ovat riippumattomia, niiden kovarianssi on nolla.

Muuttujan kovarianssi itsensä kanssa on muuttujan varianssi:

$$\operatorname{cov}(X, X) = \sigma^2 X.$$

Kovarianssin avulla voidaan määritellä korrelaatiokerroin R:

$$R(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sigma X \, \sigma Y} = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{cov}(X, X) \operatorname{cov}(Y, Y)}}.$$

Jos muuttujia on usempia, voimme laskea kaikkien muuttujaparien väliset kovarianssit. Niistä muodostettua matriisi

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix},$$

missä

$$\sigma_{ij} = \operatorname{cov}(X_i, X_j)$$

on kovarianssimatriisi. Tämä on symmetrinen matriisi, ja sen lävistäjällä olevat alkiot antavat muuttujien varianssit.