

Insinöörimatematiikka 3

Mika Hirvensalo
`mikhirve@utu.fi`

Matematiikan ja tilastotieteen laitos
Turun yliopisto

2022

Määritelmä

Jatkuvalla satunnaismuuttujalla on *standardoitu* normaalijakauma, jos sen tiheysfunktio on muotoa

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

Tällöin kertymäfunktio on muotoa

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}t^2} dt.$$

Normaalijakaumaa kutsutaan myös *Gaussin jakaumaksi* ja sen tiheysfunktion kuvaajaa *Gaussin kellokäyräksi*.

Yleinen normaalijakauma

Jakaumalla $N(\mu, \sigma^2)$ tarkoitetaan jakaumaa, jonka tiheysfunktio on

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

ja tästä johdettava kertymäfunktio

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Jakaumalle $N(\mu, \sigma^2)$ on $\mathbb{E}(X) = \mu$ ja $\text{Var}(X) = \sigma^2$.

Markovin epäyhtälö

Jos $\mathbb{P}(X < 0) = 0$ ja $a > 0$, on

$$\mathbb{P}(X > a) \leq \frac{\mathbb{E}(X)}{a}.$$

Chebyschevin epäyhtälö

$$\mathbb{P}(|X - \mathbb{E}(X)| > a) \leq \frac{\text{Var}(X)}{a^2}$$

Suurten lukujen laki

Jos $X_1, \dots, X_n$ ovat samoin jakautuneita $((\mu, \sigma^2))$, riippumattomia satunnaismuuttujia ja

$$\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$$

näiden aritmeettinen keskiarvo (ns. otoskeskiarvo), on jokaista $\epsilon > 0$ kohti

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \xrightarrow{n \rightarrow \infty} 0.$$

Keskeinen raja-arvolause

Olkoot X_i ja $\bar{X}_n$ kuten edellisessä lauseessa. Suurilla n :n arvoilla $\bar{X}_n$ jakautuu likimain normaalijakauman $N(\mu, \frac{\sigma^2}{n})$ mukaisesti.

Toisin sanoen,

$$\mathbb{P}(\bar{X}_n \leq x) \approx \Phi\left(\frac{x - \mu}{\sigma/\sqrt{n}}\right)$$

kun n on suuri.

Esimerkki

Alueelle suunnitellaan asuntoja tuhannelle perheelle. Perheissä on 0, 1, 2 tai ≥ 3 lasta todennäköisyyksillä 0,35, 0,45, 0,15 ja 0,05. Mikä on todennäköisyys sille, että 1000 koulupaikkaa ei riitä alueen lapsille?

Olkoon X_i lapsiluku perheessä i . Tällöin $\mathbb{E}(X_i) = 0,9$ (laske) ja $\text{Var}(X_i) = 0,69$ (laske). Jos merkitään $L = X_1 + \dots + X_{1000}$, on L (kokonaislapsiluku) satunnaismuuttuja, jonka odotusarvo on $0,9 \cdot 1000 = 900$ ja varianssi 690.

Näin ollen

$$\begin{aligned}\mathbb{P}(L > 1000) &= 1 - \mathbb{P}(L \leq 1000) \approx 1 - \Phi\left(\frac{1000 - 900}{\sqrt{690}}\right) \\ &= 0,000703499 \dots\end{aligned}$$

Esimerkki

Erään kaupungissa huomattiin että 344 kaikista 747:stä kuolemantapauksesta oli sattunut korkeintaan 3 kk syntymäpäivän jälkeen. Mikä on tällaisen todennäköisyys?

Jos kuolemat jakautuisivat tasaisesti, pitäisi näitä olla 3 kk syntymäpäivän jälkeen olla noin $747 \cdot \frac{1}{4} = 186,75$. 344 on kuitenkin liki 1,85-kertainen tähän nähden.

Määritetään todennäköisyys $\mathbb{P}(X \geq 344)$ sillä otaksumalla, että $p = 0,25$ on todennäköisyys kuolla minä hyvänsä 3 kuukauden ajanjaksona.

Binomijakauma

Binomijakauman perustella voidaan laskea

$$\mathbb{P}(X \geq 344) = \sum_{k=344}^{747} \binom{747}{k} 0,25^k \cdot 0,75^{747-k}.$$

Hankala laskea. Varianssi $np(1-p) = 747 \cdot 0.25 \cdot 0.75 = 140,06$, joten keskeinen raja-arvolause antaa

$$\begin{aligned}\mathbb{P}(X \geq 344) &= 1 - \mathbb{P}(0 < X < 344) \\ &\approx 1 - \left(\Phi\left(\frac{343,5 - 186,75}{\sqrt{140,06}}\right) - \Phi\left(\frac{-0,5 - 186,75}{\sqrt{140,06}}\right) \right) \\ &= 1,09491 \cdot 10^{-56} = 0,00000000 \dots 000109491 \dots\end{aligned}$$

Vrt. Suoraan binomijakauman perusteella laskettu arvo on $1,19478 \cdot 10^{-35}$.

Ongelmanasettelu

- Kaksi satunnaismuuttujaa X ja Y .
- Näiden havainnot $\mathbf{x} = (x_1, x_2, \dots, x_n)$ ja $\mathbf{y} = (y_1, y_2, \dots, y_n)$ tapauksissa $1, 2, \dots, n$.
- Selvitettävä, kuinka yleisesti $y_i \approx c \cdot x_i + b$.

Muuttujien siirto

- Merkitään $\mathbf{1} = (1, 1, \dots, 1)$
- $\mu_{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n x_i$ ja $\mu_{\mathbf{y}} = \frac{1}{n} \sum_{i=1}^n y_i$
- Määritellään $\mathbf{x}' = \mathbf{x} - \mu_{\mathbf{x}} \mathbf{1}$ ja $\mathbf{y}' = \mathbf{y} - \mu_{\mathbf{y}} \mathbf{1}$.
- Tällöin $\mathbf{x} = c\mathbf{y} + b\mathbf{1} \Leftrightarrow \mathbf{x}' = c\mathbf{y}'$.
- Lisäksi vektoreiden $\mathbf{x}'$ ja $\mathbf{y}'$ koordinaattien summa on 0

Riippuvuus

- Yhdensuuntaisuus $\mathbf{x}' = c\mathbf{y}'$ merkitsee täydellistä (lineaarista) riippuvuutta.
- Vektoreiden $\mathbf{x}'$ ja $\mathbf{y}'$ kohtisuoruus merkitsee täydellistä (lineaarista) riippumattomuutta.
- Riippuvuuden määrää on siksi luonteva mitata vektoreiden $\mathbf{x}'$ ja $\mathbf{y}'$ välisellä kulmalla (tai sen kosinilla)
-

$$\text{Corr}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}' \cdot \mathbf{y}'}{||\mathbf{x}'|| ||\mathbf{y}'||}$$

Käsitteitä

- Lähtökohta: Satunnaismuuttujat $X_1, \dots, X_n$.
- Tavoite: Satunnaismuuttujat $Y_1, \dots, Y_n$,

$$Y_i = \sum_{k=1}^n B_{ik} X_k,$$

jotka selittäisivät satunnaisvaihtelua "paremmin" kuin alkuperäiset havaintoihin perustuvat muuttujat

- Lisätavoite: Satunnaismuuttujat $Y_1, \dots, Y_n$ eivät korreloi
- Lisätavoite: Vain muutama muuttuja $Y_1, \dots, Y_l$ ($l < n$) selittäisi *suurimman* osan vaihtelusta

Käsitteitä

- $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{im})$ datavektori, jossa m havaintoa satunnaismuuttujasta X_i
- Määritelmä: $\mu_i = \mathbb{E}(X_i) = \frac{1}{m} \sum_{k=1}^m x_{ik}$.
- Määritelmä: $\mathbf{x}'_i = \mathbf{x}_i - \mu_i \mathbf{1}$.
- Määritelmä: $\text{Cov}(\mathbf{x}_i, \mathbf{x}_j) = \mathbb{E}((\mathbf{x}_i - \mu_i)(\mathbf{x}_j - \mu_j)) = \frac{1}{m} \mathbf{x}'_i \cdot \mathbf{x}'_j$
- Huom: $\text{Cov}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\|\mathbf{x}'_i\| \cdot \|\mathbf{x}'_j\|}{m} \text{Corr}(\mathbf{x}_i, \mathbf{x}_j)$
- Määritelmä: Kovarianssimatriisi

$$\Sigma = \frac{1}{m} \begin{pmatrix} \mathbf{x}'_1 \cdot \mathbf{x}'_1 & \mathbf{x}'_1 \cdot \mathbf{x}'_2 & \dots & \mathbf{x}'_1 \cdot \mathbf{x}'_n \\ \mathbf{x}'_2 \cdot \mathbf{x}'_1 & \mathbf{x}'_2 \cdot \mathbf{x}'_2 & \dots & \mathbf{x}'_2 \cdot \mathbf{x}'_n \\ \dots & \dots & \ddots & \dots \\ \mathbf{x}'_n \cdot \mathbf{x}'_1 & \mathbf{x}'_n \cdot \mathbf{x}'_2 & \dots & \mathbf{x}'_n \cdot \mathbf{x}'_n \end{pmatrix}$$

Perustuloksia

- Kovarianssimatriisin Σ ominaisarvot ovat aina reaalisia ja ≥ 0
- Kovarianssimatriisi Σ voidaan diagonalisoida: $\Sigma = B^{-1}\Lambda B$
- Matriisi B voidaan koostaa matriisin Σ ominaisvektoreista.