MTTTP5, luento 15.11.2018

Luku 4
Satunnaisotos, otossuure ja otosjakauma

4.1. Satunnaisotos

$X_1, X_2, \ldots, X_n$ on satunnaisotos, jos X_i:t ovat riippumattomia ja noudattavat samaa jakaumaa.

Sanonta "$X_1, X_2, \ldots, X_n$ on satunnaisotos $N(\mu, \sigma^2)$:sta" tarkoittaa, että jokainen $X_i \sim N(\mu, \sigma^2)$ ja X_i:t ovat riippumattomia.

4.2. Otossuureet ja otosjakaumat

- Otossuure

  satunnaisotoksen avulla määritelty funktio

- Otosjakauma

  otossuureen todennäköisyysjakauma

Otossuureita ja niiden jakaumia

1) Olkoon $X_1, X_2, \ldots, X_n$ on satunnaisotos $N(\mu, \sigma^2)$:sta. Tällöin

$$\boxed{\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).}$$

2) Olkoon $X_1, X_2, \ldots, X_n$ on satunnaisotos jakaumasta, jonka odotusarvo μ ja varianssi σ^2.

Tällöin

$$\boxed{\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \textit{likimain}.}$$

Esim. Erään tilastotoimiston (The National Center for Health Statistics) mukaan väestössä keski-ikäisten miesten verenpaineen keskiarvo on 128 ja keskihajonta 15. Haluttiin selvittää, poikkeaako keski-ikäisten yritysjohtajien verenpaineen keskiarvo koko väestön vastaavasta keskiarvosta. Mitattiin 72 yritysjohtajan verenpaineet ja saatiin keskiarvoksi 130,5. Onko eroja?

Olkoon X = verenpaine. Nyt

$$\overline{X} \sim N\left(128, \frac{15^2}{72}\right), likimain,$$

jos otos koko väestöstä.

$$P\left(\overline{X} \geq 130{,}5\right) = 1 - P\left(\overline{X} \leq 130{,}5\right) \approx 1 - \Phi\left(\frac{130{,}5 - 128}{\frac{15}{\sqrt{72}}}\right)$$

$$= 1 - \Phi(1{,}41) = 1 - 0{,}9207 = 0{,}0793$$

Ei voida ajatella, että yritysjohtajien verenpaineen keskiarvo olisi korkeampi kuin koko väestön, koska ei ole koko väestöstä tehdyssä 72 alkion otoksessa harvinaista saada otoskeskiarvoa, joka on yli yritysjohtajilta mitatun.

3) Olkoon $X_1, X_2, \ldots, X_n$ on satunnaisotos populaatiosta, jossa π % viallisia. Määritellään $X_i = \begin{cases} 1, & jos\ viallinen \\ 0, & muulloin \end{cases}$

Viallisten kokonaislukumäärä otoksessa

$X = X_1 + X_2 + \ldots + X_n \sim \mathrm{Bin}(n, \pi/100)$

Lisäksi

$$\boxed{X \overset{likimain}{\sim} N\left(\frac{n\pi}{100}, n\left(\frac{\pi}{100}\right)\left(1 - \frac{\pi}{100}\right)\right), kun\ n\ suuri.}$$

Viallisten prosenttiosuus otoksessa p = 100X/n.

E(p) = π, Var(p) = $\pi(100-\pi)/n$, ks. esim. 5.1.1.

Koska X:n jakauma on likimain normaalijakauma, niin

$$\boxed{p \overset{likimain}{\sim} N\left(\pi, \frac{\pi(100-\pi)}{n}\right), kun\ n\ suuri.}$$

Esim. 4.2.3 Olet todistamassa oikeudessa, jossa väitetään erään pelipaikan ruletin toimivan väärin. Ruletissa on 37 numeroa, joiden kaikkien pitäisi olla yhtä todennäköisiä. Pelipaikka voittaa numerolla nolla. Olet saanut selville, että 3700 kertaa rulettia pyöritettäessä nolla tuli 140 kertaa. Millaisen todistuksen annat oikeudessa?

Olkoon X = nollien lukumäärä.

Jos ruletti toimii oikein, niin X ~ Bin(3700, 1/37).

E(X) = 3700·(1/37) = 100,

Var(X) = 3700·(1/37)·(36/37) = 3600/37.

Tällöin X ~ N(100, 3600/37), likimain.

$P(X \geq 140) = 1 - P(X \leq 139) \approx 1 - \Phi\left(\frac{139-100}{\sqrt{3600/37}}\right)$

$= 1 - \Phi(3{,}95) \approx 0$. Tämä on siis lähes mahdotonta.

Todistan, että pelipaikan ruletti toimii väärin.

Esim. Yritys tekee tiettyä komponenttia, jota käytetään auton moottorissa. Tämä komponentti hajoaa joskus heti, kun se on otettu käyttöön. Yritys valvoo tuotantoaan siten, että virheellisten komponenttien osuus ei saisi olla suurempi kuin 4 %. Laaduntarkkailussa tehtiin 500 komponentin otos, joista 28 komponenttia osoittautui virheelliseksi. Onko tuotanto keskeytettävä?

Ratkaisu 1

Olkoon X = virheellisten komponenttien lukumäärä 500 alkion otoksessa.

Jos tuotannossa virheellisiä 4 %, niin

$X \sim \text{Bin}(500,\ 0{,}04)$, jolloin $E(X) = 500 \cdot 0{,}04 = 20$,

$\text{Var}(X) = 500 \cdot 0{,}04 \cdot 0{,}96 = 19{,}2$.

Lisäksi $X \sim N(20,\ 19{,}2)$, likimain.

$$
\begin{aligned}
P(X \geq 28) &= 1 - P(X \leq 27) \\
&\approx 1 - \Phi((27 - 20)/\sqrt{19{,}2}) \\
&= 1 - \Phi(1{,}60) = 0{,}0548.
\end{aligned}
$$

Tämä ei harvinaista, tuotantoa voidaan jatkaa.

Binomijakaumasta laskettuna $P(X \geq 28) = 0{,}0489$,

ks. http://homepage.stat.uiowa.edu/~mbognar/applets/bin.html

Ratkaisu 2

Olkoon p = virheellisten komponenttien prosenttiosuus 500 alkion otoksessa

Jos tuotannossa virheellisiä 4 %, niin

$$p \overset{likimain}{\sim} N(4,\ \underbrace{\frac{4(100-4)}{500}}_{0{,}768})$$

P(p ≥ 5,6) = 1 – P(p ≤ 5,6)

≈ 1 – Φ((5,6 – 4)/√0,768)

= 1 – Φ(1,83) = 1 – 0,9664 = 0,0336

Sama tulos ratkaisusta 1, jos lasketaan

$P(X \geq 28) \approx 1 - \Phi((28 - 20)/\sqrt{19{,}2}) = 1 - \Phi(1{,}83)$.

4) Olkoot $X_1, X_2, \ldots, X_n$ satunnaisotos $N(\mu_1,\ \sigma_1^2)$:sta ja $Y_1, Y_2, \ldots, Y_m$ satunnaisotos $N(\mu_2,\ \sigma_2^2)$:sta sekä otokset riippumattomia.

Tällöin

$$\boxed{\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)}$$

Esim. Tarkastellaan lapsen syntymäpainoa grammoina. Oletetaan, että tytöillä syntymäpaino $X \sim N(3450, 520^2)$ ja pojilla syntymäpaino $Y \sim N(3640, 440^2)$. Tarkastellaan tyttöpopulaatiosta 100 alkion ja poikapopulaatiosta 200 alkion satunnaisotoksia. Määritä otoskeskiarvojen jakaumat sekä otoskeskiarvojen erotuksen jakauma. Laske todennäköisyys sille, että tyttöjen otoskeskiarvo on suurempi kuin poikien.

$$\overline{X} \sim N\left(3450, \frac{520^2}{100}\right)$$

$$\overline{Y} \sim N\left(3640, \frac{440^2}{200}\right)$$

$$\overline{X} - \overline{Y} \sim N\left(3450 - 3640, \frac{100^2}{100} + \frac{440^2}{200}\right)$$

$$\overline{X} - \overline{Y} \sim N(-190, 3672), joten\ P\left(\overline{X} - \overline{Y} > 0\right)$$

$$= 1 - \Phi\left(\frac{0 - (-190)}{60,6}\right) = 1 - \Phi(3,14)$$

$$= 0,0008$$

Esim. 4.2.4 Koneiden A ja B pitäisi valmistaa keskimäärin samanmittaisia tankoja. Molempien koneiden tuottamien tankojen pituuksissa X ja Y (cm) on jonkin verran vaihtelua, jota voidaan luonnehtia normaalijakaumalla, jonka varianssi on 0,20 cm^2. Laadunvalvonnassa seurataan koneiden toimintaa ja valitaan satunnaisesti koneen A tuotannosta 20 ja koneen B tuotannosta 10 tankoa. Koneen A tuotannosta valittujen tankojen keskipituus on 40,0 cm ja koneelta B valittujen 39,5 cm. Tuottavatko koneet keskimäärin samanmittaisia tankoja?

$$\overline{X} - \overline{Y} \sim N\left(0, \underbrace{\frac{0{,}2}{20} + \frac{0{,}2}{10}}_{0{,}03}\right), jos\ tuottavat\ samanmittaisia$$

$$P(\left|\overline{X} - \overline{Y}\right| > 0{,}5 = 1 - P\left(-0{,}5 \le \overline{X} - \overline{Y} \le 0{,}5\right)$$

$$= 1 - \left(\Phi\left(\frac{0{,}5 - 0}{\sqrt{0{,}03}}\right) - \Phi\left(\frac{-0{,}5 - 0}{\sqrt{0{,}03}}\right)\right)$$

$$= 1 - \left(\Phi(2{,}89) - \Phi(-2{,}89)\right) = 2 - 2\Phi(2{,}89)$$

$$= 0{,}0038$$

Eivät tuota keskimäärin samanmittaisia tankoja. Jos tuottaisivat, niin olisi harvinaista saada otokset, joiden keskiarvojen erotuksen itseisarvo olisi suurempi kuin 0,5 cm.