Tilastomatematiikka

Ensimmäinen välikoe 18.02.2021

1. Laske todennäköisyys $\mathbb{P}(X \geq 1)$, kun

   a) $X \sim \text{Bin}(200, 0.01)$.

   b) $X \sim N(2, 1.98)$.

   c) $X \sim \text{Poi}(2)$.

2. Tieto siitä, kuka on sairastanut koronan, voidaan selvittää vasta-ainetestien avulla. Vasta-ainetestin tulokseen liittyy epävarmuutta seuraavista syistä johtuen.

   (i) Koronan sairastaneen testitulos on positiivinen todennäköisyydellä 93.8 %. (*sensitiivisyys*)

   (ii) Testitulos on negatiivinen todennäköisyydellä 95.6 % henkilölle, joka ei ole sairastanut koronaa. (*spesifisyys*)

   Oletetaan, että Oulun väkiluku on 208000 ja että koronan olisi sairastanut 4000 oululaista. Oletetaan lisäksi, että kaikille oululaisille tehtäisiin nyt vasta-ainetesti.

   a) Kuinka monen oululaisen vasta-ainetestin tulos on positiivinen? (2p)

   b) Millä todennäköisyydellä positiivisen testituloksen saanut on sairastanut koronan? Piirrä tilannetta havainnollistava puukaavio. (4p)

3. Suositussa *Pokemon Go*-pelissä jahdataan niin sanotuissa *raideissa* toinen toistaan komeampia legendaarisia pokemoneja, joita pelipiireissä kutsutaan *legeiksi*. Legellä on 3 yksilöllistä ominaisuutta X_1, X_2, X_3, jotka voivat toisistaan riippumatta olla kokonaislukuja $10, 11, 12, 13, 14, 15$, jotka arvotaan täysin satunnaisesti. Yksilön *laadun prosentteina* ilmoittaa satunnaismuuttuja

$$Y = \frac{100}{45}\left(X_1 + X_2 + X_3\right).$$

   a) Laske muuttujan X_i, $i = 1, 2, 3$, odotusarvo. (2p)

   b) Laske muuttujan X_i, $i = 1, 2, 3$, varianssi. (2p)

   c) Laske normaalijakauma-approksimaatiolla todennäköisyys, että raidista saadaan vähintään 90-prosenttinen lege. Mitä voit sanoa approksimaation tarkkuudesta, kun simuloimalla saatu kolmidesimaalinen likiarvo on 0.162? (4p)

Kaavoja

Todennäköisyyden ominaisuuksia

$$\begin{aligned}
&\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B),\\
&\mathbb{P}(A \backslash B) = \mathbb{P}(A \cap \overline{B}) = \mathbb{P}(A) - \mathbb{P}(A \cap B),\\
&\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A),\\
&\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B),\\
&\mathbb{P}(B|A) = \frac{\mathbb{P}(B)\mathbb{P}(A|B)}{\mathbb{P}(A)}
\end{aligned}$$

Odotusarvoja ja variansseja

Ptnf. tai tf.	$\mu_X := \mathbb{E}(X)$	$\sigma_X^2 := \mathrm{Var}(X)$
$\mathbb{P}(X = x)$	$\sum_x x\mathbb{P}(X = x)$	$\sum_x (x - \mu_X)^2 \mathbb{P}(X = x)$
$f_X(x)$	$\int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x$	$\int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)\mathrm{d}x$
$\binom{n}{x} p^x (1-p)^{n-x}$	np	$np(1-p)$
$p(1-p)^{x-1}$	$1/p$	$(1-p)/p^2$
$\frac{a^x}{x!}\mathrm{e}^{-a}$	a	a
$1/(b-a)$	$(a+b)/2$	$(b-a)^2/12$
$\theta \mathrm{e}^{-\theta x}$	$1/\theta$	$1/\theta^2$
$\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y), \quad \mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

Eräitä testimuuttujia

$$\begin{aligned}
&\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \text{ (likimain, kun "n on suuri")},\\
&\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},\\
&\frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{1}{n} + \frac{1}{m}}\sqrt{\frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}}} \sim t_{n+m-2},\\
&\sqrt{n-1}s_x \frac{\frac{S_{xY}}{s_{xx}} - \beta}{S_r} \sim t_{n-2}
\end{aligned}$$

Regressio, korrelaatio ja kovarianssi

$$\begin{aligned}
&r = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}};\ s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y});\ s_{xx} = s_x^2;\\
&y = a + bx; \quad b = \frac{s_{xy}}{s_{xx}}; \quad a = \overline{y} - b\overline{x};\\
&s_r^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = \frac{n-1}{n-2}(1 - r^2)s_{yy};\\
&\sigma_{XY} = \mathrm{Cov}(X,Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))), \quad \sigma_{XX} = \sigma_X^2;\\
&\qquad = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)\\
&\rho(X,Y) = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}
\end{aligned}$$

Tehtävien ratkaisuperiaatteet

1. a) Koska X saa vain ei-negatiivisia kokonaislukuarvoja, kannattaa laskea todennäköisyys ”komplementin kautta”

$$\mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0) = 1 - 0.99^{200} \approx 0.87.$$

b) Käytetään hyväksi standardointia $Z = \frac{X - \mathbb{E}(X)}{\sqrt{\mathrm{Var}(X)}} = \frac{X-2}{\sqrt{1.98}} \sim N(0,1)$, jonka mukaan

$$\begin{aligned}
\mathbb{P}(X \geq 1) &= \mathbb{P}\left(\frac{X-2}{\sqrt{1.98}} \geq \frac{1-2}{\sqrt{1.98}}\right) \approx \mathbb{P}(Z \geq -0.71) && \text{(standardointi)} \\
&= 1 - \Phi(-0.71) = \Phi(0.71) && \text{(symmetriaominaisuus)} \\
&\approx 0.76. && \text{(taulukosta)}
\end{aligned}$$

c) Menetellään tässä samalla tavalla kuin a)-kohdassa, jolloin kaavakokoelmasta saatavan Poisson-jakauman pistetodennäköisyyden laskukaavalla saadaan

$$\mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0) = 1 - \mathrm{e}^{-2} \approx 0.86.$$

2. Merkitään

$$\begin{aligned}
S &= \text{”henkilö on sairastanut koronan”}, \\
T &= \text{”testitulos on positiivinen”}.
\end{aligned}$$

Tiedetään todennäköisyydet $\mathbb{P}(S) = \frac{4000}{208000} = \frac{1}{52}$, $\mathbb{P}(T|S) = 0.938$ ja $\mathbb{P}(\overline{T}|\overline{S}) = 0.956$. Viimeisimmästä todennäköisyydestä saadaan $\mathbb{P}(T|\overline{S}) = 1 - \mathbb{P}(\overline{T}|\overline{S}) = 0.044$.

a) Virheellisistä diagnooseista johtuen positiivisia tulee myös ei-sairastaneista. Toisaalta testi ei tavoita kaikkia sairastaneita. Sairastaneista tulee $0.938 \cdot 4000 = 3752$ ja ei-sairastaneista $0.044 \cdot (208000 - 4000) = 8976$ positiivista näytettä. Yhteensä nämä tekevät 12728 positiivista näytettä.

b) Kysytään todennäköisyyttä $\mathbb{P}(S|T)$, joka Bayesin kaavan mukaan on

$$\mathbb{P}(S|T) = \frac{\mathbb{P}(T|S)\mathbb{P}(S)}{\mathbb{P}(T)}.$$

Osoittajan todennäköisyydet tunnetaan, mutta nimittäjän todennäköisyys on tuntematon. Se taas saadaan kokonaistodennäköisyyden kaavalla

$$\mathbb{P}(T) = \mathbb{P}(T|S)\mathbb{P}(S) + \mathbb{P}(T|\overline{S})\mathbb{P}(\overline{S}) = 0.938 \cdot \frac{1}{52} + 0.044 \cdot \left(1 - \frac{1}{52}\right).$$

Edellisen perusteella kysytty todennäköisyys on

$$\mathbb{P}(S|T) = \frac{0.938 \cdot \frac{1}{52}}{(0.938 + 51 \cdot 0.44)\frac{1}{52}} = \frac{938}{3182} \approx 29\%.$$

Samaan lopputulokseen päästään luonnollisesti puukaaviolla ja a)-kohdasta saaduilla luvuilla, joiden mukaan todennäköisyys on $\frac{3752}{12728}$. Jätetään puukaavion piirtäminen harjoitustehtäväksi. ”Puun rungosta” lähtevä haara kannattaa tehdä sairastamisen mukaan, sillä sairastaminen on ehtona näytteiden lopputuloksille (ehdollisille todennäköisyyksille).

3. a) Koska eri vaihtoehtoja yksilöllisille ominaisuuksille on 6, on kunkin arvon todennäköisyys $\frac{1}{6}$. Odotusarvoksi saadaan

$$\mu_i = \mathbb{E}(X_i) = \frac{1}{6}(10 + 11 + 12 + 13 + 14 + 15) = 12.5.$$

b) Varianssi on määritelmän mukaan

$$\sigma_i^2 = \mathrm{Var}(X_i) = \sum_{k=10}^{15} (k - \mu_i)^2 \mathbb{P}(X_i = k) = 2.916667.$$

c) Merkitään $S_3 = X_1 + X_2 + X_3$, jolloin $Y = \frac{100}{45} S_3$. Kysytään todennäköisyyttä $\mathbb{P}(Y \geq 90) = \mathbb{P}\left(S_3 \geq \frac{90 \cdot 45}{100}\right) = \mathbb{P}(S_3 \geq 40.5)$.
Lasketaan summan odotusarvo ja varianssi. Riippumattomuuden nojalla

$$\mathbb{E}(S_3) = \mu_1 + \mu_2 + \mu_3 = 3 \cdot 12.5 = 37.5 \quad \text{ja} \quad \mathrm{Var}(S_3) = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 \approx 8.75.$$

Samalla tavalla kuin tehtävässä 1 saadaan

$$\begin{aligned}
\mathbb{P}(S_3 \geq 40.5) &= \mathbb{P}\left(\frac{S_3 - \mathbb{E}(S_3)}{\sqrt{\mathrm{Var}(S_3)}} \geq \frac{40.5 - 37.5}{\sqrt{8.75}}\right) && \text{(standardointi)} \\
&\approx 1 - \Phi(1.01) && \text{(normaaliapproksimaatio)} \\
&\approx 1 - 0.8438 \approx 0.156. && \text{(taulukosta)}
\end{aligned}$$

Tulosta voi pitää varsin hyvänä ottaen huomioon, että yhteenlaskettavia on vain 3.