2B Odotusarvot

Tuntitehtävät

2B1 (Potenssin odotusarvo) Satunnaismuuttujalla X on jatkuva jakauma tiheysfunktiolla

$$f(x) \;=\; \begin{cases} 1, & 0 < x < 1, \\ 0, & \text{muuten.} \end{cases}$$

Olkoon n jokin positiivinen kokonaisluku.

(a) Määritä satunnaismuuttujan $Y = X^n$ kertymäfunktio ja tiheysfunktio.
Vihje: Mitä arvoja X^n voi saada? Kertymäfunktio $F_Y(y)$ ilmaisee erään tapahtuman todennäköisyyden. Minkä tapahtuman, ja milloin kyseinen tapahtuma toteutuu?

(b) Laske $\mathrm{E}(X^n)$ käyttäen (a)-kohdassa määritettyä X^n:n tiheysfunktiota.

(c) Laske $\mathrm{E}(X^n)$ käyttäen odotusarvon muunnoskaavaa (luento 2A / monisteen luku 3.3).

(d) Käyttäen johtamaasi kaavaa, laske $\mathrm{E}(X^n)$ kun $n = 1, 2, 3, 4$.

Ratkaisu.

(a) Satunnaismuuttuja X^n saa arvoja välillä $(0, 1)$. Kaikilla $0 < t < 1$ pätee

$$\mathrm{P}(X^n \le t) \;=\; \mathrm{P}(X \le t^{1/n}) \;=\; \int_0^{t^{1/n}} 1 \, ds \;=\; t^{1/n}.$$

Satunnaismuuttujan $Y = X^n$ kertymäfunktio on siis

$$F_Y(t) \;=\; \begin{cases} 0, & t \le 0, \\ t^{1/n}, & 0 < t \le 1, \\ 1, & t > 1. \end{cases}$$

Derivoimalla saadaan tiheysfunktioksi

$$f_Y(t) \;=\; \begin{cases} (1/n)\, t^{1/n-1}, & 0 < t < 1, \\ 0, & \text{muuten.} \end{cases}$$

(b) Kysytty odotusarvo saadaan kaavasta

$$\mathrm{E}(Y) \;=\; \int_{-\infty}^{\infty} t f_Y(t)\, dt \;=\; (1/n) \int_0^1 t^{1/n}\, dt \;=\; (1/n) \Big/_{t=0}^{1} \left(\frac{t^{(1/n)+1}}{(1/n)+1} \right) \;=\; \frac{1}{1+n}.$$

(c) Muunnosfunktio on $g(x) = x^n$, niin että $Y = g(X)$. Odotusarvon muunnoskaavalla

$$\mathrm{E}(X^n) \;=\; \int_{-\infty}^{\infty} t^n f(t)\,dt \;=\; \int_0^1 t^n\,dt \;=\; \frac{1}{1+n}.$$

Tulos on sama kuin b-kohdassa, mutta lasku oli helpompi.

(d) $\mathrm{E}(X^1) = 1/2$ (kuten pitäisikin, koska $X^1 = X$)
$\mathrm{E}(X^2) = 1/3$
$\mathrm{E}(X^3) = 1/4$
$\mathrm{E}(X^4) = 1/5$.

Vapaaehtoinen lisäharjoitus: Arvo 10000 satunnaislukua välillä $(0, 1)$, esim. R:n komennolla `runif` tai Matlabin/Octaven komennolla `unifrnd`. Korota ne n:nteen potenssiin ja tutki potenssien keskiarvoa (komennolla `mean`). Vertaa *havaittua keskiarvoa* matemaattisesti laskettuun *odotusarvoon*.

2B2 (Odotusaikaparadoksi) Bussit saapuvat pysäkillesi tiettyinä aikoina, kolme bussia tunnissa. Saavut pysäkille X minuuttia yli 9, missä X on tasajakautunut avoimella välillä $]0, 60[$ (ts. välin päätepisteet eivät ole mukana).

(a) Jos bussit saapuvat säännöllisesti 20 minuutin välein, mikä on odotusaikasi odotusarvo?

(b) Luettuasi aikataulun huomaat, että bussit saapuvatkin epätasaisin mutta säännöllisin välein kello 9:00, 9:10, 9:30, 10:00, ... jne. Esitä bussin odotusaika w oman saapumisaikasi x funktiona $w = g(x)$ ja piirrä funktio. (Vihje: Määrittele funktio paloittain.)

(c) Laske $\mathrm{E}(W)$, missä $W = g(X)$. Vihje: Odotusarvon muunnoskaava.

(d) Vertaa kohtien (a) ja (c) tuloksia ja selitä arkijärjellä.

Ratkaisu.

(a) 10 minuuttia.

(b) Jaetaan väli $]0, 60[$ kolmeen palaan (osaväliin). Kullakin osavälillä pätee, että jos seuraava bussi saapuu hetkellä b (yli yhdeksän), niin odotusaikasi on $b - x$, joka tietysti riippuu x:n arvosta.

$$g(x) = \begin{cases} 10 - x & \text{jos } 0 < x \le 10 \\ 30 - x & \text{jos } 10 < x \le 30 \\ 60 - x & \text{jos } 30 < x < 60. \end{cases}$$

(Funktion kuvaaja on sahalaitainen käyrä.)

(c) Lasketaan integraali paloittain. X:n tiheysfunktio on vakio $f(x) = \frac{1}{60}$ koko välillä $]0, 60[$.

$$\begin{aligned}
\mathrm{E}(W) &= \int_0^{60} g(x) f(x)\, dx = \frac{1}{60} \int_0^{60} g(x)\, dx \\
&= \frac{1}{60} \int_0^{10} (10 - x)\, dx \;+\; \frac{1}{60} \int_{10}^{30} (30 - x)\, dx \;+\; \frac{1}{60} \int_{30}^{60} (60 - x)\, dx \\
&\approx 0.8333 + 3.3333 + 7.5000 = \mathbf{11.6666}.
\end{aligned}$$

Integroinnin sijasta *voisit* päätellä seuraavasti. Todennäköisyytesi saapua pysäkille näillä kolmella osavälillä ovat 1/6, 2/6 ja 3/6. *Jos* saavut ensimmäisellä välillä, niin odotusaikasi odotusarvo on 5 minuuttia; vastaavasti 10 ja 15 minuuttia 2. ja 3. osavälillä. Sitten lasketaan painotettu keskiarvo $\frac{1}{6} \cdot 5 + \frac{2}{6} \cdot 10 + \frac{3}{6} \cdot 15 \approx 11.667$ minuuttia. Tähän päättelyyn kuitenkin tarvitaan ns. *ehdollisen odotusarvon* käsite. Se, että koko W:n odotusarvo saadaan ehdollisten odotusarvojen painotettuna keskiarvona eri tapauksista, perustuu *iteroidun odotusarvon lauseeseen* (vrt. todennäköisyyden osituskaava, luento 1B). Näiden osaamista ei vaadita tällä kurssilla, mutta niillä lasku sujuisi helpommin.

(d) C-kohdassa odotusajan odotusarvo on pidempi. Tämä johtuu siitä, että henkilö saapuu todennäköisemmin pitkän bussivälin aikana kuin lyhyen bussivälin aikana.

Tämä *odotusaikaparadoksi* (engl. waiting time paradox) on eräs tapaus yleisempää *kokovinoumaa* (size bias), jossa jostakin suureesta otettu otos tai näyte on "vinoutunut" koska suureen arvo vaikuttaa otantaprosessiin.

Bussipysäkillä voisit pitää kirjaa kunakin päivänä toteutuneesta bussin odotusajasta. Kun pitkän ajan kuluessa toteaisit, että joudut odottamaan keskimäärin $11\frac{2}{3}$ minuuttia, saattaisit päätellä, että bussit kulkevat siis keskimäärin $23\frac{1}{3}$ minuutin välein. Kuitenkin ne kulkevat 3 kertaa tunnissa eli keskimäärin 20 minuutin välein. Oma arviosi vuorovälistä olisi *ylöspäin vinoutunut* koska et havaitse eri vuorovälejä tasaisin todennäköisyyksin $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ vaan havaintosi painottuvat pidempiin vuoroväleihin.

Reaalimaailman esimerkkejä kokovinoumasta löytyy monilta tieteenaloilta. Esimerkki tekstiiliteollisuudesta: Yrität arvioida kuitujen keskipituutta poimimalla suuren määrän yksittäisiä kuituja koneellisilla pihdeillä, ja laskemalla *poimittujen* kuitujen keskipituuden. Mutta pihdit poimivat pitkiä kuituja todennäköisemmin kuin lyhyitä kuituja, joten poimitut kuidut ovatkin vinoutunut otos. Lisätietoa ilmiöstä esim. Arratia, Goldstein & Kochman: "Size bias for one and all", `https://arxiv.org/abs/1308.2729`.

Kotitehtävät

2B3 (Epidemia) Eräässä epidemiassa sairaiden lukumäärän arvioidaan R-kertaistuvan joka viikko. Kasvuvauhtia kuvaava kerroin R on tuntematon luku, mutta sama luku joka viikko. Epidemiologi Adam arvioi, että R on tasajakautunut välillä $I = [0.6, 1.4]$, ts. hän pitää yhtä todennäköisenä, että R on esim. välillä $[0.60, 0.61]$ kuin millä tahansa muulla samanpituisella välillä, joka sisältyy I:hin. Kerroin $Y = R^{12}$ ilmaisee monikokertaiseksi sairaiden määrä kasvaa tai pienenee 12 viikossa.

(a) Laske $\mathrm{E}(Y)$.

(b) Millä välillä ovat Y:n mahdolliset arvot?

(c) Adamin kaveri Bertil arvelee, että Y on tasajakautunut (b)-kohdassa lasketulla välillä. Laske, mikä on kyseisen tasajakauman odotusarvo ja päättele tästä, voiko Bertil olla oikeassa.

(d) Toinen kaveri Cecil arvioi, että Y:n odotusarvo on $(\mathrm{E}(R))^{12}$. Laske kyseinen luku ja kerro mitä mieltä olet Cecilin arviosta.

Kohdista (e1) ja (e2) riittää **jommankumman** tekeminen. Molemmatkin saa tehdä, jos intoa riittää!

(e1) Selvitä Y:n kertymäfunktio (vihje: luento 2A), sen perusteella Y:n tiheysfunktio, ja piirrä tiheysfunktion kuvaaja alueella $y > 0.5$. Kuvaile Y:n jakaumaa sanallisesti. Laske myös todennäköisyydet $\mathrm{P}(Y > 1)$ ja $\mathrm{P}(Y > 10)$.

(e2) Tutki Y:n jakaumaa kokeellisesti tietokoneella seuraavasti. Arvo 100 000 mahdollista R:n arvoa (vihje: Matlabissa/Octavessa `unifrnd` tai R:ssä `runif`), laske vastaavat Y:n arvot näissä tapauksissa, ja piirrä niistä histogrammi (vihje: Matlab/Octave/R `hist`). Kuvaile Y:n jakaumaa sanallisesti. Laske myös *suhteelliset esiintyvyydet* tapahtumille $\{Y > 1\}$ ja $\{Y > 10\}$.

Arviointiohje. 0.4 pistettä per kohta. Kokonaistulos pyöristetään ylöspäin. Jos on tehnyt sekä e1- että e2-kohdat, niistäkin saa kummastakin 0.4 pistettä (kuitenkin koko tehtävästä enintään 2 pistettä).

Ratkaisu.

(a) R:n tiheysfunktio on $f_R(r) = 1.25$, kun $0.6 \le r \le 1.4$. Odotusarvon muunnoskaavan mukaisesti

$$\mathrm{E}(Y) = \mathrm{E}(R^{12}) = \int_{0.6}^{1.4} r^{12} f_R(r)\, dr = 1.25 \cdot \int_{0.6}^{1.4} r^{12}\, dr = \frac{1.25}{13} \Big/ {}_{0.6}^{1.4} (r)^{13} \approx \mathbf{7.65}.$$

Odotusarvomielessä sairaiden määrän siis arvioidaan kasvavan noin 7.65-kertaiseksi.

(b) Koska $0.6 \le R \le 1.4$, niin $0.6^{12} \le R^{12} \le 1.4^{12}$ eli $Y \in [a, b] \approx [\mathbf{0.002}, \mathbf{56.694}]$.

(c) Kyseisen välin tasajakauman odotusarvo olisi $(a+b)/2 \approx (0.002 + 56.694)/2 \approx \mathbf{28.348}$. Koska a-kohdassa laskettiin, että Y:n odotusarvo on jotain aivan muuta, niin selvästikään Y ei voi olla tasajakautunut kyseisellä välillä, eli Bertil on jo tämän perusteella väärässä. (Näemme asian tarkemmin e-kohdassa, jossa selvitämme Y:n jakauman.)

(d) Cecilin arvio on $(\mathrm{E}(R))^{12} = 1^{12} = 1$. Tämä on aika pahasti pielessä. Satunnaismuuttujan muunnoksen R^{12} odotusarvo *ei ole* sama kuin kyseinen muunnosfunktio sovellettuna R:n odotusarvoon.

(e1) Jos y on jokin piste välillä $[a, b]$ (ks. b-kohta), niin

$$F_Y(y) = \mathrm{P}(Y \le y) = \mathrm{P}(R^{12} \le y) = \mathrm{P}(R \le y^{1/12}) = \frac{y^{1/12} - 0.6}{1.4 - 0.6} = 1.25 \cdot (y^{1/12} - 0.6).$$

Tästä derivoimalla

$$f_Y(y) = \frac{5}{48} \cdot y^{-11/12}.$$

Y:llä on selvästi epätasainen jakauma: tiheys laskee kohti mahdollisen välin oikeaa päätepistettä.

Kertymäfunktiosta ja komplementtisäännöstä saadaan suoraan

$$\mathrm{P}(Y > 1) = 1 - F_Y(1) = 1 - 1.25 \cdot (1^{1/12} - 0.6) = \mathbf{0.5}$$
$$\mathrm{P}(Y > 10) = 1 - F_Y(10) = 1 - 1.25 \cdot (10^{1/12} - 0.6) \approx \mathbf{0.236}$$

(e2) Histogrammin pitäisi olla selvästi oikealle laskeva, kuten tiheysfunktionkin. Suhteelliset esiintyvyydet voi laskea esim. Matlabissa/Octavessa komennolla `sum(Y>10)/n`. Suhteellisten esiintyvyyksien pitäisi olla suunnilleen samat kuin e1-kohdassa laskettujen todennäköisyyksien.

2B4 (Vertaisarviointi) Laskuharjoitukseen saapuu 20 opiskelijaa, joista jokainen palauttaa kotitehtävien vastauspaperinsa assistentille. Assistentti sekoittaa vastauspaperit huolellisesti ja jakaa ne sitten takaisin opiskelijoille tarkastettaviksi, yhden kullekin. Määritä odotusarvo niiden opiskelijoiden lukumäärälle, jotka päätyvät tarkastamaan oman vastauspaperinsa.

Vihje. Määritellään indikaattorimuuttuja

$$X_i = \begin{cases} 1, & \text{jos } i\text{:s opiskelija tarkastaa oman vastauspaperinsa,} \\ 0, & \text{muuten.} \end{cases}$$

Odotusarvon lineaarisuudesta voi myös olla apua.

Arviointiohje. Pisteitä saa seuraavista havainnoista:

- Oman vastauspaperinsa tarkastavien opiskelijoiden lukumäärän esittäminen indikaattorien summana
- Odotusarvon lineaarisuuden käyttäminen
- Indikaattorimuuttujan odotusarvon laskeminen: $\mathrm{E}(X_i) = \mathrm{P}(X_i = 1)$
- Tapahtuman "i:s opiskelija arvioi oman vastauspaperinsa" todennäköisyyden laskeminen.

Jokaisesta yo. havainnosta saa 0.5 pistettä. Tehtävän kokonaispisteet pyöristetään ylöspäin lähimpään kokonaislukuun.

Ratkaisu. Oman vastauspaperinsa tarkastavien opiskelijoiden lukumäärä on satunnaismuuttuja, joka voidaan kirjoittaa muodossa

$$X \;=\; X_1 + \cdots + X_{20},$$

missä X_i on kuten vihjeessä. Odotusarvon lineaarisuuden perusteella

$$\mathrm{E}(X) \;=\; \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_{20}).$$

Koska indikaattorimuuttuja X_i saa vain arvoja 0 ja 1, on sen odotusarvo

$$\begin{aligned} \mathrm{E}(X_i) \;&=\; 0 \times \mathrm{P}(X_i = 0) + 1 \times \mathrm{P}(X_i = 1) \\ &=\; \mathrm{P}(X_i = 1). \end{aligned}$$

Pitää siis vielä laskea tn, jolla i:s opiskelija arvioi oman vastauspaperinsa. Koska vastauspaperit jaetaan tasaisen satunnaisesti kaikkien opiskelijoiden kesken, tämä todennäköisyys on $\frac{1}{20}$ (opiskelijan saamaan paperiin on 20 vaihtoehtoa, joista vain yksi on suotuisa). Kysytty odotusarvo on siis

$$\mathrm{E}(X) \;=\; \sum_{i=1}^{20} \mathrm{P}(X_i = 1) \;=\; 20 \times \frac{1}{20} \;=\; 1.$$

Huom: Yllä laskettu odotusarvo ei riipu opiskelijoiden lukumäärästä.

Lineaarisuuden avulla saimme X:n *odotusarvon* laskettua helposti. Emme kuitenkaan selvittäneet X:n *jakaumaa*. Sekin on mahdollista tehdä, mutta selvästi työläämpää. Tiedämme tietysti, että X saa kokonaislukuarvoja joukossa $\{0, 1, \ldots, 20\}$, mutta millä todennäköisyyksillä? Ks. esim. Wikipedia: Rencontres numbers, `https://en.wikipedia.org/wiki/Rencontres_numbers`.

Tämä on tyypillistä todennäköisyyslaskennassa: On usein helpompaa laskea jostakin suureesta *vain* odotusarvo eikä koko jakaumaa tarkasti. Näin saadaan suureen jakaumasta edes *jotain* tietoa.

Yleisen tapauksen sijasta voi ajatella ainakin pieniä erikoistapauksia. Jos opiskelijoita on vain 1, hän saa varmasti oman paperinsa ($X = 1$ varmasti). Jos taas opiskelijoita on 2, niin yhtä todennäköisesti he saavat kumpikin oman paperinsa ($X = 2$) tai kumpikin toisen paperin ($X = 0$). Ainakin näissä tapauksissa selvästi nähdään, että $\mathrm{E}(X) = 1$ pätee. Entäs jos opiskelijoita on kolme . . . ?