MS-A0501

TODENNÄKÖISYYSLASKENNAN JA TILASTOTIETEEN PERUSKURSSI. 4.9.–16.10.2023. Aalto-yliopisto. Yliopistonlehtori Pekka Pere.

Koe 16.10.2023

Koe alkaa 9.00 ja päättyy 12.00.

Esitä kaikkien laskujesi välivaiheet, ja perustele kaikki vastauksesi yksityiskohtaisesti. Pelkkä oikea vastaus on nollan pisteen arvoinen. Kaikki tehtävät ovat kuuden pisteen arvoisia. Parhainta koemenestystä!

1. Suomessa ajettiin 1900-luvun alkupuolella moottoripyörällä kilpaa niin, että maaliin tarkimmin etukäteen määrätyssä ajassa tullut voitti. Ajan saavuttaminen vaati kovaa vauhtia, mutta radan nopeimmin ajanut ei siis välttämättä voittanut. Kilpailunjärjestäjä mittasi kilpailijoiden maaliintuloajat kronometrillä (erityisen tarkalla kellolla). Kilpailijat sopeuttivat vauhtinsa omien kellojensa mukaan. Tehtävänlaatijan ystävä kertoi isoisänsä voittaneen monia tällaisia kilpailuja, koska hän mittasi ajoaikaansa kahden oman kellonsa mittaaman ajan keskiarvolla. Muut kilpailijat käyttivät vain yhtä kelloa. Kerran isoisä ei voittanut kilpailua. Tällöin hän kyseenalaisti kilpailunjärjestäjän kronometrin tarkkuuden. Tarkkuusmittauksessa isoisän kellojen keskiarvo osoittautui kilpailunjärjestäjän kronometriä tarkemmaksi. Pohditaan selityksiä ja kahden kellon käytön edun suuruutta.

a) Oletetaan, että sekä kilpailijoiden että isoisän molempien kellojen poikkeamat järjestäjän etukäteen osoittamasta tavoiteajasta maaliintulolle noudattavat toisistaan riippumattomasti normaalijakaumaa $\mathsf{N}(0, \sigma^2)$, jossa $\sigma^2 > 0$. Mitä jakaumaa noudattaa isoisän kellojen mittaamien poikkeamien keskiarvo? Perustele huolellisesti sanoin ja kaavoin jakauma ja miksi isoisä sai etua kahden kellon käytöstä.

b) Oletetaan, että kilpailunjärjestäjän kronometrin mittaama tavoiteaika poikkeaa tarkistusmittauksen mukaisesta todellisesta ajasta normaalijakauman $\mathsf{N}(0, \sigma^{*2})$ mukaisesti. Kuinka suuri voi σ^2 olla, jotta isoisän kellojen mittaamien aikojen keskiarvo voi vielä olla tarkempi ajan mittari kuin kronometri? Vastaa perustelulla ja kaavalla.

c) Olkoon $\sigma^{*2} = 4$. Kuinka suuri voi isoisän yhden kellon mittaaman ajan keskihajonta σ olla, jotta kellojensa mittaamien aikojen keskiarvo voi olla tarkempi kuin kronometri? Kuinka suuri on isoisän yhden kellon mittaaman ajan varianssi tällöin? Vastaa perustelulla ja lukuarvoilla.

2. Avaruusluotaimessa on tietojen tallentamista varten neljä kovalevyä. Avaruuden ankarissa oloissa kullakin levyllä on todennäköisyys 0.1 hajota ennenaikaisesti (luotaimen suunniteltuna toiminta-aikana), toisista levyistä riippumatta.

a) Levyt yhdistetään RAID0-järjestelmäksi, joka lakkaa toimimasta jos yksikin levy hajoaa. Mikä on todennäköisyys, että järjestelmä lakkaa toimimasta? Vastaus ainakin neljällä desimaalilla.

b) Levyt yhdistetään RAID6-järjestelmäksi, joka toimii kunhan ainakin kaksi levyistä toimii. Mikä on todennäköisyys, että järjestelmä lakkaa toimimasta? Vastaus ainakin neljällä desimaalilla.

3. Olkoot satunnaismuuttujat X_i riippumattomia ja olkoon niillä sama odotusarvo $\mathsf{E}(X_i) = \mu$ ja varianssi $\mathsf{V}(X_i) = \sigma^2$, $i = 1, \ldots, n$. Havaintojen lukumäärän kasvaessa kohti ääretöntä todennäköisyys, että keskiarvo

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

poikkeaa μ:stä vähemmän kuin mielivaltaisen pienen positiivisen vakion ($\epsilon > 0$) verran, lähestyy yhtä. Tulosta kutsutaan suurten lukujen laiksi. Todista tulos. (Vihje1: Tshebyshevin/Tšebyšovin epäyhtälö $\mathsf{P}(|X - \mu| \leq k\sigma) \geq 1 - 1/k^2$ kaikilla $k \geq 1$. Vihje2: Valitse $k = \epsilon/(\sigma/\sqrt{n})$.)

4. Luennoijalla on neljä 6-sivuista noppaa, joista kolme on tavallisia (tulokset $1, \ldots, 6$ ovat yhtä todennäköiset). Yksi noppa on litistynyt siten, että tulosten 1 ja 6 todennäköisyys on kummankin 0.3 ja kunkin muun tuloksen todennäköisyys 0.1. Luennoija on poiminut yhden nopista satunnaisesti ja heittänyt sitä viisi kertaa saaden tulosjonon $(1, 6, 6, 3, 1)$. Merkitään $\Theta = 1$ jos poimittu noppa on tavallinen ja $\Theta = 2$ jos se on litistynyt.

a) Laske todennäköisyys saada juuri tämä tulosjono, jos noppa oli tavallinen.

b) Laske todennäköisyys saada juuri tämä tulosjono, jos noppa oli litistynyt.

c) Määritä parametrin Θ posteriorijakauma.

d) Käyttäen saatua posteriorijakaumaa, määritä todennäköisyys saada 10 kuutosta, jos poimittua noppaa heitetään vielä 10 kertaa.

5.

a) Floridan valtionyliopisto erotti kriminologian professori Eric Stewartin 2023. Syy oli epäselvyydet hänen julkaisemissaan tutkimuksissa. Kuusi hänen arvostetuissa aikakauskirjoissa julkaisemaa tutkimusta on mitätöity (*retracted*). Pickett (2020) kritisoi Stewartin tutkimuksia.[1] Yksi Pickettin kritiikeistä on pähkinänkuoressa, että Stewart on raportoinut Bernoulli-jakautuneen satunnaismuuttujan keskihajontoja, jotka eivät pidä paikkaansa. Jos esimerkiksi tapahtuman todennäköisyys on 0.86, niin satunnaismuuttujan keskihajonnan tulisi Pickettin mielestä olla 0.35 eikä Stewartin laskema 0.41. Perustele väite sanoin ja kaavoin.

b) Abeysooriya ym. (2021) tutkivat 11 117 tieteellistä artikkelia, joissa esiintyi geenien nimiä ja joiden aineistot olivat Excel-muodossa.[2] Excel-tiedostoista 3 436:ssa eli noin 30.9 %:ssa ainakin yhden geenin nimi oli tallentunut väärin. Esimerkkinä Excel-tiedostoista löytyneistä vääristä geenien nimistä tutkijat osoittavat nimen "Jan-41", jonka on ilmeisesti ollut tarkoitus olla TAMM41. Tutkijat arvelevat virheen syntyneen siitä, että Excel on tulkinnut, että "TAMM" tarkoittaa suomen kielen tammikuuta.

Laske 95 %:n kaksisuuntainen luottamusväli osuudelle tieteellisiä artikkeleita, joiden aineisto on Excel-tiedostossa ja geenin nimi on muuttunut tiedostossa. Voit olettaa, että Abeysooriya ym:iden aineisto on satunnaisotos. (Vihje: Standardinormaalijakauman 0.975. kvantiili on 1.960 (`qnorm(0.975)`).

---

[1] J.T. Pickett (2020): The Stewart Retractions: A Quantitative and Qualitative Analysis. *Econ Journal Watch*, 17, 152–190.

[2] M. Abeysooriya, M. Soria, M. S. Kasu ja M. Ziemann (2021): Gene Name Errors: Lessons not Learned. *PLOS Computational Biology*, 17(7): e1008984. `https://doi.org/10.1371/journal.pcbi.1008984`.