JOHDATUS TODENNÄKÖISYYSLASKENTAAN JA TILASTOLLISEEN PÄÄTTELYYN. 1.1.–26.2.2021. Tampereen yliopisto. Yliopistonlehtori Pekka Pere.

Koe 26.2.2021

Koe alkaa 17.00 ja päättyy 20.00.

Esitä kaikkien laskujesi välivaiheet, ja perustele kaikki vastauksesi yksityiskohtaisesti. Pelkkä oikea vastaus on nollan pisteen arvoinen. Kaikki tehtävät ovat kuuden pisteen arvoisia. **Palauta vastauksesi neljään kysymykseen. Riippumatta siitä, kuinka monta tehtävää olet palauttanut, vastaa Moodlessa 6. kysymykseen, mitkä tehtävät tulee arvostella.** Jos et vastaa, arvostelu ei välttämättä perustu parhaisiin tai jättämiisi vastauksiisi.

Vastausten tulee olla käsinkirjoitettuja. Myös lyhyet R-koodit, jos sellaisia esitätte vastauksissanne, tulee olla käsinkirjoitettuja. Vastauksissa tulee käyttää kaavoja ja perustella laskut ja päätelmät huolella. (Esim. pelkkä R-käsky ja oikea vastaus eivät ole riittävä selitys.)

Skannatkaa kunkin tehtävän vastauksenne pdf-tiedostoksi ja palauttakaa se Moodlessa kyseisen tehtävän kohtaan. Lähettäkää vastauksia kysymyksiin pitkin koeaikaa. Jos jättäisitte kaikkien vastausten jättämisen viime tinkaan ja kaikki tekisivät niin, Moodle saattaisi tukkeutua ettekä saisi lähetettyä vastauksianne. Ennen jättämistä tarkistakaa vastauksenne huolellisesti. Ja vielä toisen kerran!

Kaikissa pdf-tiedostoissa tulee heti alussa näkyä selkeästi nimenne ja opiskelijanumeronne. Tarkistakaa ennen vastauksen lähettämistä, että se on tallentunut kokonaisena ja selkeänä. Huom! Skannaussovellus saattaa laittaa pdf:ään logon tai vastaavan. Se ei saa peittää vastauksenne osaa. Tehtävästä saa pisteitä vain vastauksen mukaan, joka näkyy pdf-tiedostossa.

Voitte käyttää Moodle-sivun luentomonistetta ja taulukoita, mitä tahansa tilastotieteellistä kirjallisuutta, Googlea, Internettiä, kuinka tahansa kehittyneitä laskimia, tietokonetta, tilasto-ohjelmia jne. apuna vastaamisessanne. Toista ihmistä ette saa millään tavalla käyttää apunanne. Ette saa pohtia vastauksia tai vastata ryhmissä, soittaa neuvoja, s-postitse, tekstiviestitse tai millään muulla tavalla olla kokeen aikana kehenkään yhteydessä, joka voisi auttaa teitä vastauksissa.

Saatan liittää kokeeseen suullisen kuulustelun joillekin opiskelijoille, jos hahmotan sen heidän osaltaan tarpeelliseksi. Suullisen kuulustelun läpäisy on heille edellytys tentin läpäisemiselle. Ilmoitan tällaisesta tarpeesta kyseisille opiskelijoille kokeiden tarkastamisen jälkeen.

Tarpeen vaatiessa lähetän teille s-postitse lisäohjeita kokeen aikana. Teidän tulee palauttaa vastauksenne ennen koeajan loppumista. Vakavassa ongelmatilanteessa minulle voi s-postittaa (pekka.pere@tuni.fi) tai soittaa (050 437 7568).

Koemenestystä!

1. Pohditaan, voitaisiinko ryhmätestauksella tehostaa (nopeuttaa ja halventaa) viruksen kantajuuden testaamista. Ihmisiä tulee virustestiin terveysasemalle. Kukin on tai ei ole viruksen kantaja toisistaan riippumattomasti. Jaetaan virusnäytteet 30 ryhmiin. Ilmeinen menettely 1: Tutkitaan kaikki 30 näytettä kussakin ryhmässä. Mahdollisesti nopeampi ja halvempi menettely 2: Jaetaan ryhmän 30 näytettä kahteen osaan, muodostetaan 30 näytteen puolikkaista yksi yhdistetty näyte ja tutkitaan se. Jos yhdistetyssä näytteessä ei ole virusta, kenessäkään ryhmäläisessä ei ole virusta ja kaikki ryhmäläiset julistetaan viruksettomaksi. On selvitty yhdellä kokeella. Jos yhdistetystä näytteestä löytyy virus, tutkitaan loput 30 näytteen puolikasta ja arvioidaan erikseen kunkin ryhmäläisen kantajuus. On jouduttu tekemään 31 koetta 30 sijaan. Näin jatketaan kaikkien ryhmien kohdalla.

Lääkintäviranomainen haluaa käyttää menettelyä, joka johtaa keskimäärin pienempään laboratoriokoemäärään. Jos viruksen kantajuuden todennäköisyys on suuri, 30:n ryhmässä tapaa olla aina ainakin yksi kantaja, ja menettely 1 johtaa pienempään määrään kokeita. Selvitetään, kuinka pieni pitää viruksen kantajuuden todennäköisyyden olla, jotta viranomaisen kannattaa käyttää menettelyä 2.[1]

Sovelletaan menetelmää 2.

a) Oletetaan, että viruksen kantajuuden todennäköisyys on π. Mikä on todennäköisyys, että tarvitaan vain yksi testi?

b) Mikä on todennäköisyys, että tarvitaan 31 testiä?

c) Mikä on testien lukumäärän odotusarvo ylipäänsä? Entä jos $\pi = 0$ tai $\pi = 1$?

d) Kannattaako menetelmää 2 käyttää, jos $\pi = 0.001$, $\pi = 0.01$ tai $\pi = 0.1$? (Vihje: R:llä voi laskea potensseja `x^30`-tapaisilla komennoilla.)

2. Anna Soudakova kertoo esikoisteoksessaan (2020) Mitä männyt näkevät isoisovanhemmistaan. He olivat muuttaneet Suomesta Neuvostoliittoon (1922–1991) rakentamaan kommunistista yhteiskuntaa. Heidät pidätettiin 1936 isänmaan vihollisina ja ilmeisesti teloitettiin 1937 Sandarmohissa Venäjän Karjalassa. Joukkohaudat löytänyt Juri Dmitrijev tuomittiin 2020 kolmeksitoista vuodeksi vankilaan Venäjällä. Soudakovaa huolettaa, että Venäjällä kohdellaan taas ihmisiä isänmaan vihollisina. Tehtävä alla on vuodelta 1969 muttei täysin vanhentunut.[2]

Oletetaan, että olet poliittinen vanki Neuvostoliitossa ja sinut aiotaan karkoittaa joko Siperiaan tai Mongoliaan. Todennäköisyys joutua Siperiaan on 0.7 ja Mongoliaan 0.3. Lisäksi tiedetään, että valittaessa umpimähkään [satunnaisotannalla] siperialainen hän todennäköisyydellä 0.8 käyttää hylkeennahkaturkkia; vastaava todennäköisyys Mongoliassa on 0.4. Eräänä iltamyöhänä silmäsi sidotaan ja sinut heitetään kuorma-auton lavalle. Kaksi viikkoa myöhemmin (arviointisi mukaan) kuorma-auto pysähtyy, sinulle kerrotaan, että olet perillä karkotuspaikassasi ja silmäside poistetaan. Ensimmäisellä henkilöllä, jonka näet [satunnaisotantaa], ei ole yllään hylkeennahkaturkkia. Mikä on todennäköisyys, että karkoituspaikkasi on Siperia[?].

---

[1] Tehtävä on muunnelma Larsenin ja Marxin (2001, 197) esimerkistä.

[2] Lähteitä: `https://fi.wikipedia.org/wiki/Juri_Dmitrijev` (haettu 24.2.2021). P.-M. Vasama ja Y. Vartia (1980): *Johdatus tilastotieteeseen — Osa I.* 4. painos. Gaudeamus. S. 328.

3. Kauppias myy vuosittain 1 200 kovalevyä. Niistä keskimäärin 12 palautetaan kauppiaalle viallisina. Oletetaan, että viallisten uusien kovalevyjen lukumäärä (Y) noudattaa Poisson-jakaumaa.

a) Estimoi Poisson-jakauman parametri μ. Oletetaan, että viallisten kovalevyjen lukumäärä noudattaa Poisson-jakaumaa estimoidulla parametrilla. Määrittele viallisten kovalevyjen lukumäärän kertymäfunktio. Mitkä ovat jakauman odotusarvo ja varianssi?

b) Mikä on todennäköisyys, että vuotuinen viallisten kovalevyjen lukumäärä on korkeintaan 12? Vähintään 19? Onko jakauma symmetrinen? Perustele.

4. HPV eli papilloomavirus aiheuttaa naisten syövistä 9 % ja miesten syövistä 1 %:n. HPV-rokote antaa lähes 100 %:n suojan HPV:n aiheuttamalta kohdunkaulansyövän esiasteelta. (Lehtinen ym. 2018.[3]) HPV tarttuu yleensä seksuaalisessa kanssakäymisessä. Ilman rokotetta 80–90 % aikuisista saa HPV-tartunnan elämänsä aikana; toisaalta 90 % tartunnan saaneista paranee kahdessa vuodessa. Vuodesta 2013 lähtien 11–12-vuotiaita tyttöjä on rokotettu HPV-rokotteella. Mahdollisuuden ottaa rokote on hyödyntänyt toivottua pienempi osuus 70 %. Terveyden ja hyvinvoinnin laitos (THL) suositteli 2019, että pojat otettaisiin mukaan rokotusohjelmaan, koska HPV aiheuttaa miehillekin syöpiä ja jotta laumasuoja syntyisi. (Kotaniemi-Talonen ym. 2019.[4]) THL laajensi rokotusohjelman koskemaan poikia syksyllä 2020.[5]

Tehtävä koskee HPV-tartunnan yleisyyttä ennen rokotusten alkamista. Auvinen ym. (2004, 2005)[6] keräsivät tutkimuksiinsa kaksi otosta Ylioppilaiden terveydenhoitosäätiön asiakkaista pääkaupunkiseudulla. Ensimmäinen otos koostuu terveystarkastukseen 2001–2003 tulleista ensimmäisen vuoden naisyliopisto-opiskelijoista ($n = 919$). Toinen otos on ehkäisyvälinemääräystä lääkäriltä hakeneita oireettomia naisyliopisto-opiskelijoita 2002–2004 ($n = 550$). Otokset koostuvat opiskelijoista, jotka antoivat suostumuksensa osallistua tutkimukseen. Yhdistetyssä otoksessa HPV-kantajia on 33 % (taulukko). Auvinen ym. (2005) pitävät osuutta suurena ja mahdollisesti merkityksellisenä syöpien ilmaantumisen kannalta.

	HPV-kantaja	ei-HPV-kantaja	%-kantajia
otos 1	310	609	33.7
otos 2	175	375	31.8
otokset 1 ja 2	485	984	33.0

a) Laske 99 %:n luottamusväli HPV-kantajille ensimmäisen vuoden naisyliopisto-opiskelijoiden keskuudessa pääkaupunkiseudulla. Kattaako luottamusväli kantajaosuuden 50 %?

b) Voiko tehtävässä annettujen tietojen perusteella otosten olettaa olevan satunnaisotoksia vastaavista populaatioista? Mitkä ovat kyseiset populaatiot? Mi-

[3] M. Lehtinen, P. Nieminen ja J. Paavonen (2018): HPV-rokotuksen vaikuttavuus Suomessa. *Duodecim*, 134, 1281–1228.

[4] L. Kotaniemi-Talonen, M. Jakobsson, A. Virtanen ja P. Nieminen (2019): HPV ja kohdunkaulasyövän ehkäisy — missä meillä nyt mennään? *Duodecim*, 135, 1889–1897.

[5] `https://thl.fi/fi/web/infektiotaudit-ja-rokotukset/rokotteet-a-o/hpv-eli-papilloomavirusrokote/poikien-hpv-rokotukset` (haettu 3.2.2021).

[6] E. Auvinen, M. Niemi, C. Malm, R. Zilliacus, A. Trontti, R. Fingerroos, M. Lehtinen ja J. Paavonen (2005): High Prevalence of HPV among Female Students in Finland. *Scandinavian Journal of Infectious Diseases*, 37, 873–876.

tä päättelet? (Vihje: Aineiston keruutavassa on yhteisiä piirteitä HIV-esimerkin kanssa.)[7]

5. Taulukko on Rowenin ja Emeryn (2018) artikkelista.[8] Havainnot ovat pistemääriä, jotka kuvaavat, kuinka paljon vanhempi mustamaalaa (*denigrate*) toista vanhempaa (enemmän mustamaalaamista; suurempi pistemäärä). Rivin "eronneet" kukin havainto on saatu yhdeltä lapselta, joka on kertonut tiedot molemmista vanhemmistaan (326 lasta). Rivin "naimisissa" tiedot on saatu vastaavasti kunkin pariskunnan lapsista yhdeltä (668 lasta). Taulukon soluissa on pistemäärien keskiarvot ($\hat{\mu}_{ij}$) ja keskihajonnat (s_{ij}). Oikeanpuolimmaisessa sarekkeessa on t-testisuureet ja niiden p-arvot testeistä, joissa verrataan vastaavan rivin odotusarvoja. Alimmalla rivillä on t-testisuureet ja niiden p-arvot testeistä, joissa verrataan vastaavan sarakkeen odotusarvoja. Oletetaan, että pistemäärät ovat normaalijakautuneita.

	äiti	isä	
eronneet	$\hat{\mu}_{11} = 33.01$, $s_{11} = 14.12$	$\hat{\mu}_{12} = 29.20$, $s_{12} = 13.92$	$t = 5.86$, $p < 0.001$
naimisissa	$\hat{\mu}_{21} = 22.32$, $s_{21} = 9.53$	$\hat{\mu}_{22} = 20.90$, $s_{22} = 8.19$	$t = 5.18$, $p < 0.001$
	$t = 14.12$, $p < 0.001$	$t = 11.82$, $p < 0.001$	

a) Vastaa tähän kohtaan $\hat{\mu}_{ij}$-estimaattien perusteella. Mustamaalaako jompikumpi sukupuoli vanhemmista enemmän toistaan ryhmässä "naimisissa". Entä ryhmässä "eronneet"? Kummat mustamaalaavat enemmän isiä: Naimisissa olevat vai eronneet äidit? Kummat mustamaalaavat enemmän äitejä: Naimisissa olevat vai eronneet isät?

b) Ovatko p-arvot päteviä alimmalla rivillä? Entä oikeanpuolimmaisessa sarakkeessa? Perustele. Jos perustelet, etteivät ole, selitä ilmeinen (ei kuviteltavissa oleva) syy, miksi eivät ole.

c) Olkoot nolla-ja vastahypoteesit seuraavat. H_0: Eronneiden ja naimisissa olevien isien pistemäärien varianssit ovat samat. H_1: Eronneiden isien pistemäärän varianssi on suurempi. Laske testisuure ja sen p-arvo. Mitä päättelet? Onko oletus normaalijakautuneisuudesta tärkeä p-arvon tulkinnan kannalta? Perustele.

[7]Kiitän Eeva Auvista avusta tehtävän laatimisessa 2.–3.2.2021.

[8]J. Rowen ja R. Emery (2018): Parental Denigration: A Form of Conflict that Typically Backfires. *Family Court Review*, 56, 258–268.