TILASTOLLISTEN MENETELMIEN PERUSTEET. 8.3.–23.4.2021. Tampereen yliopisto. Yliopistonlehtori Pekka Pere.

Koe 23.4.2021

Koe alkaa 17.00 ja päättyy 21.00.

Esitä kaikkien laskujesi välivaiheet, ja perustele kaikki vastauksesi yksityiskohtaisesti. Pelkkä oikea vastaus on nollan pisteen arvoinen. Tehtävät ovat kuuden pisteen arvoisia paitsi ensimmäinen, jossa on ylimääräinen lisäpistekohta. Lisäpistekohdalla voit parantaa arvosanaasi; arvostelua ei alenna, jos et vastaa lisäpistekohtaan.

Vastausten tulee olla käsinkirjoitettuja. Myös lyhyet R-koodit, jos sellaisia esität, tulee olla käsinkirjoitettuja. Vastauksissa tulee käyttää kaavoja ja perustella laskut ja päätelmät huolellisesti. (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 kaikki palauttaisivat vastauksensa viime tingassa, 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. 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.

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 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). Hätätilanteessa palauttakaa vastauksenne minulle s-postitse.

Parhainta koemenestystä!

1. Marcinkowska ym. (2014) tutkivat, pitävätkö miehet naisten kasvojen feminiinistä piirteistä yhtälailla eri maissa.[1] Naisten kasvokuvia muutettiin tietokoneohjelmalla vähemmän tai enemmän feminiineiksi. Kokeeseen osallistuneet miehet 28 eri maasta arvioivat, kumpi kuvista miellyttää heitä enemmän. Kunkin maan regressiossa selitettäväksi havainnoksi (28 kappaletta) kirjattiin osuus miehiä, jotka pitivät feminiinisemmäksi muokattua kuvaa miellyttävämpänä (feminiinisyysmieltymysindeksi fm). Näitä osuuksia selitettiin regressiossa kunkin maan hyvinvointi-indeksillä (h). Se saa sitä suurempia arvoja, mitä paremmin ihmiset voivat. Aineiston maista huonoimmin voidaan Nepalissa ja parhaiten Japanissa. Suomi on neljäs pylpyrä oikealta (kuva).

Miesten testosteronitaso tapaa olla alhainen köyhissä ja korkea vauraissa maissa. Yksi teoria on, että alhaisella testosteronitasolla miehiä miellyttävät vähemmän feminiiniset ja korkealla testosteronitasolla feminiinisemmät naiset. Oletetaan, että fm- ja h-muuttujat täyttävät (likimäärin) regressiomallin perusoletukset. Tutkitaan edellä kuvattua teoriaa estimoimalla PNS:llä yhtälö

$$\begin{array}{rcccccc} fm & = & 0.683426 & + & 0.015584h & + & \hat{\varepsilon} \\ & & (0.007146) & & (0.002909) & & \end{array}$$

$$n = 28, \quad s = 0.1944, \quad \mathsf{R}^2 = 0.5247, \quad F = 28.7.$$

Luvut suluissa ovat estimoituja keskivirheitä. Mallin sovite on piirretty oheiseen kuvioon.

a) Mikä on fm:n ja h:n otoskorrelaatio?

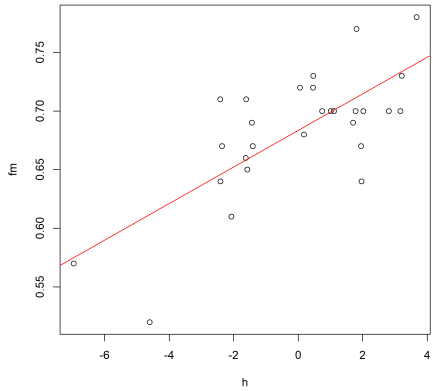
b) Testaa 1 %:n merkitsevyystasolla (kaksisuuntainen testi), onko fm:n ja h:n välillä lineaarinen yhteys. Mikä on nollahypoteesi?

c) Mitä päättelet? Tukevatko estimaatit ja testi teoriaa testosteronimäärän ja mieltymyksen feminiinisiin piirteisiin yhteydestä? Perustele.

d) Lisäpistekohta (1 p): Oletetaan, että fm-osuudet on laskettu yhtäsuuristä vastaajaotoksista miehiä kussakin maassa. Selitä huolellisesti, miksi voidaan olettaa, että fm on (likimäärin) normaalijakautunut samalla varianssilla kullakin h:n arvolla nollahypoteesin pätiessä.

Todellisuudessa vastaajaotokset olivat erisuuria eri maissa. Selitä, miksi PNS-menetelmä ei ole tällöin optimaalinen estimointimenetelmä. Mikä olisi parempi menetelmä? Perustele.

[1] U.M. Marcinkowska, M.V. Kozlov, H. Cai, J. Contreras-Garduño, B.J. Dixson, G.A. Oana, G. Kaminski, N.P. Li, M.T. Lyons, I.E. Onyishi, K. Prasai, F. Pazhoohi, P. Prokop, S.L. R. Cardozo, N. Sydney, J.C. Yong ja M.J. Rantala (2014): Cross-Cultural Variation in Men's Preference for Sexual Dimorphism in Women's Faces. *Biology Letters*, 10, 20130850. `http://dx.doi.org/10.1098/rsbl.2013.0850`. Lisää aiheesta Marcinkowskan väitöskirjassa Evolution and Sources of Individual Variation in Mate Preferences in Humans. Turun yliopiston julkaisuja 290. Kuva on piirretty artikkelin dataliitteen aineistosta. Sen luvut on ilmeisesti pyöristetty alkuperäisistä. Tehtävän kuvan havainnot eroavat hieman ko. artikkelin ja väitöskirjan (mts. 28) kuvien havainnoista.



Kuva 1: Mieltymys feminiinisiin piirteisiin (fm) ja hyvinvointi (h).

2. Nuoren tytön itsevarmuus (jatkoa). Mallitetaan nuoren tytön itsevarmuutta osoitinmuuttujilla äidin (x) ja isän (z) kasvatustyyleistä (laiminlyövä = 1, autoritaarinen = 2, salliva = 3 ja auktoritatiivinen = 4). Osoitinmuuttujat saavat arvon 1 vanhemman kasvatustyylin voimassaollessa ja 0 muuten. Estimoidaan PNS-menetelmällä mallit alla (keskivirheet ovat suluissa):

$$\begin{array}{rcccccccccc} \hat{y} & = & 3.252 & - & 0.414x_1 & - & 0.135x_4 & + & 0.131z_4 & + & 0.548x_4z_4. \\ & & (0.122) & & (0.191) & & (0.174) & & (0.265) & & (0.311) \end{array}$$

$$s = 0.796,\ \mathsf{R}^2 = 0.1614,\ F = 8.275,\ n = 177.$$

ja

$$\begin{array}{rcccccc} \hat{y} & = & 3.206 & - & 0.364x_1 & + & 0.590x_4z_4. \\ & & (0.082) & & (0.171) & & (0.134) \end{array}$$

$$s = 0.794,\ \mathsf{R}^2 = 0.1557,\ F = 16.040,\ n = 177.$$

a) Mikä suure R^2 on? Miksi se on pienempi alemmassa mallissa? Mikä suure s on? Miksi se on pienempi alemmassa mallissa? (Vihje: Pohdi, miten s^2 määritellään ja miksi se voi käyttäytyä tehtävässä esitetyllä tavalla.)

b) Mitä hypoteesia F-testisuure testaa? Mikä on sen p-arvo alemmassa mallissa? Mitä päättelet sen perusteella?

c) Testaa alemman mallin yksittäisten regressiokertoimien estimaattien tilastollista merkitsevyyttä merkitsevyystasolla 0.05 (kaksisuuntaiset testit). Mitä päättelet?

d) Tulkitse alempi malli huolellisesti (vakion ja regressiokertoimien estimaatit ja mihin verrataan, minkä arvon itsevarmuus kussakin tilanteessa saa).

3. Lapsen psyykkisen hyvinvoinnin yhteys perherakenteeseen oli väitöksen aihe.[2] Väitöstiedotteesta 29.11.2013:

> Tuore Itä-Suomen yliopistossa tarkastettava väitöstutkimus osoittaa, että uusperheiden[3] lapsilla oli enemmän psyykkisiä ongelmia kuin yksinhuoltajaperheiden lapsilla tai niiden perheiden lapsilla, joissa oli kaksi biologista vanhempaa. – – Uusperheiden lapsilla käyttäytymisen ongelmat yleisempiä. – – tutkimuksessa – – havaittiin, että uusperheiden lasten riski psyykkiselle oireilulle vaikuttaisi suuremmalta kuin yksinhuoltajaperheiden lasten.

Väitöskirjassa todetaan (sivut 6, 72, 92 ja 96):

> Tarkastelin perheeseen liittyvien tekijöiden, muuttujien ja lastenpsyykkisten häiriöiden yhteyksiä ensin erikseen, lopuksi vein vielä muuttujat binaariseen logistiseen regressioanalyysiin, jotta havaitsisin ovatko kaikki kyseiset asiat lapsen psyykkisen voinnin kannalta tilastollisesti merkitseviä. – – Yksittäisinä tekijöinä tutkittuna perherakenne, dynamiikka ja arvot olivat yhteydessä lapsen psyykkiseen vointiin. Nämä muuttujat vietiin tutkimuksessa vielä binaariseen logistiseen regressioanalyysiin, jotta voitiin tarkastella, olivatko ne edelleen kaikki merkittäviä – – . – – **yksinhuoltajuus ei näyttäisi olevan suurempi riskitekijä kuin uusperheisyys. – – uusperheiden lapset näyttäisivät riskiytyvän psyykkiselle oireilulle vielä enemmän kuin yksinhuoltajaperheiden lapset.**

Summeeraukset edellä perustuvat ilmeisesti ainakin osin väitöskirjan taulukkoihin 9 ja 21. Tehtävän kannalta selventäviä tietoja taulukosta 9 “Lasten psyykkiset häiriöt ja perherakenne – – ” on alla (n on havaintojen lukumäärä kussakin ryhmässä).

perhemuoto	lapsilla psyykkisiä ongelmia lukumäärä (%)	n
biologinen	75 (12.69)	591
uusperhe	17 (30.91)	55
yksinhuoltaja	23 (21.10)	109
yhteensä	115 (15.23)	755

Taulukko 21 on “Logistinen regressioanalyysi lapsen psyykkisiin ongelmiin liittyville tekijöille”. Tehtävän kannalta oleelliset tiedot siitä ovat alla. (OR on ristisuhde eli *odds ratio*. Suluissa on kyseisen osoitinmuuttujan luokat.)

[2] R. Väänänen (2013): Perheen rakenteen, dynamiikan ja arvojen merkitys lapsen psyykkiselle hyvinvoinnille. Väitöskirja. Yhteiskuntatieteiden ja kauppatieteiden tiedekunta. Itä-Suomen yliopisto. `http://urn.fi/URN:ISBN:978-952-61-1271-8`. Tiedote väitöskirjasta: `http://www.uef.fi/fi/uef/-/29-11-toimiva-vuorovaikutus-perheissa-ehkaisee-lasten-psyykkisia-hairioita` (haettu 14.12.2013). Lainauksissa ohessa on korjattu alkuperäislähteiden painovirheitä.

[3] Uusperhe on perhe, jonka eronnut vanhempi on perustanut uuden puolison kanssa.

selittävä muuttuja	OR	OR:n 95 %:n luottamusväli
perherakenne (biologiset vanhemmat/muut)	4.5	2.013–10.233
perhedynamiikan toimivuus (alhaisempi/korkeampi)	4.2	2.204–7.932
Sosiaaliset ja emotionaaliset arvot (korkeat/matalat)	2.1	1.042–4.184
sukupuoli (poika/tyttö)	4.2	2.202–8.036

a) Ovatko väitöskirjassa esitetyt johtopäätökset lasten psyykkisten ongelmien riskin erosta uusperheiden ja yksinhuoltajaperheiden välillä oikeutettuja tehdyn logistisen regression perusteella? Perustele. (Vihje: Pohdi relevantin osoitinmuuttujan luokkia.) (2 p)

b) Väitöskirjassa tulkitaan tehtävän toista taulukkoa näin (mts. 93):

> Yhteenvetona voidaan todeta, että mikäli lapsen perhe ei ole lapsen kahden biologisen vanhemman perhe, riski psyykkisille ongelmille on 4.5-kertainen. Mikäli perhedynamiikan toimivuus on keskiarvoa alhaisempi, riski psyykkisille häiriöille on 4.2-kertainen ja mikäli sosiaaliset ja emotionaaliset arvot perheessä ovat keskimääräistä alhaisemmat, riski häiriöille on kaksinkertainen. Lisäksi pojilla on tyttöihin verrattuna 4.2-kertainen riski psyykkisille häiriöille.

Miksi nämä tulkinnat ovat virheellisiä? (4 p)

4. Jääkiekon SM-liigassa pelasi 14 joukkuetta vuosina 2009 ja 2008. Joukkueiden sijoitusten yhteyttä mittaavat Spearmanin korrelaatiokerroin ja Kendallin τ ovat 0.4065934 ja 0.2967033. Testaa molempien tunnuslukujen avulla merkitsevyystasolla 0.05, voiko nollahypoteesin joukkueiden sijoitusten riippumattomuudesta hylätä (kaksi yksisuuntaista testiä; suureen otoskokoon perustuvat testit riittävät). (Vihje: $\tau = (K-D)/[n(n-1)/2]$ ja $\mathsf{V}(K-D) = n(n-1)(2n+5)/18$.)