Oulun yliopiston matemaattisten tieteiden laitos/tilastotiede

806113P TILASTOTIETEEN PERUSTEET, kl 2011 (Esa Läärä)

L-harjoitus 1, viikko 3 (to 20.1.): kotitehtävät

1. Etsi kuluvan ja mahdollisesti edeltävänkin viikon sanomalehdistä, radio- tai tv-uutisista tai niiden verkkosivuilta uutisia tai artikkeleita, joissa kerrotaan jostakin ajankohtaisesta tilastollisesta tutkimuksesta tai selvityksestä. Kuinka monta löydät ja millaisista aiheista?

Mikä löytämistäsi uutisista/kirjoituksista on itseäsi kiinnostavin juttu? Saatko selville, mikä siinä on ollut pääkysymys, kohdejoukko, havaintojen hankinnan asetelma ja mittausmenetelmät, keskeiset tulokset ja kuinka niitä on tulkittu? Mitä lisätietoja kaipaisit ollaksesi paremmin informoitu ao. tutkimuksesta ja sen tuloksista?

Todennäköisyyslaskennan kertausta

Seuraavalla sivulla annettuja kotitehtäviä 2. – 4. varten kerrataan tässä lyhyesti eräitä Todennäköisyyslaskennan peruskurssilla käsiteltyjä asioita normaali- ja binomijakauman ominaisuuksista ja niihin liittyvien todennäköisyyksien laskemisesta.

Olkoon Z satunnaismuuttuja, joka noudattaa standardinormaalijakaumaa, eli $Z \sim N(0,1)$, jonka tiheysfunktion $\varphi(z)$ ja kertymäfunktion $\Phi(z)$ lausekkeet ovat (ks. Tuominen 1993, s. 60-61)

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2), \qquad \Phi(z) = \int_{-\infty}^{z} \varphi(u)du, \quad z \in \mathbb{R}.$$

Z:n odotusarvo on 0 ja varianssi 1. Sen **p-fraktiili** z_p eli **p-kvantiili** toteuttaa ehdon (ks. Tuominen 1993, s. 90)

$$\Phi(z_p) = p, \quad \text{eli} \quad z_p = \Phi^{-1}(p), \quad p \in]0,1[.$$

Erityisiä fraktiileja ovat mm. mediaani eli 50% fraktiili $z_{0.5} = 0$ sekä alakvartiili $z_{0.25} \approx -0.675$ ja yläkvartiili $z_{0.25} \approx 0.675$.

$N(0,1)$-jakauma on symmetrinen origon suhteen; ts. kaikilla $z \in \mathbb{R}$ sekä $p \in]0,1[$ pätee:

$$\varphi(-z) = \varphi(z), \quad \Phi(-z) = 1 - \Phi(z), \quad z_p = -z_{1-p}.$$

Normaalijakaumaa $N(\mu, \sigma^2)$ noudattavan satunnaismuuttujan X, jonka odotusarvo on $\mathbb{E}(X) = \mu$ ja varianssi $D^2(X) = \sigma^2$, tiheysfunktio $f(x)$ ja kertymäfunktio $F(x)$ saadaan Z:n vastaavista funktioista (ks. Tuominen 1993, s. 62-65):

$$f(x) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right), \qquad F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right), \quad x \in \mathbb{R}.$$

ja X:n kvantiilit x_p saadaan yhtälöstä $x_p = \mu + \sigma z_p$, kun $0 < p < 1$.

Tehtäväpaperin liitteenä on kaksi taulukkoa, jotka sisältävät $N(0,1)$-jakauman tiheys- ja kertymäfunktion kuin myös fraktiilien arvoja valituilla argumenttien z ja p arvoilla (lähde: MAOL-taulukot).

Binomijakaumaa parametrein $n \in \mathbb{N}_+$ ja $p \in]0,1[$ noudattavan satunnaismuuttujan X, merk. $X \sim \text{Bin}(n,p)$, pistetodennäköisyydet noudattavat kaavaa (ks. Tuominen 1993, s. 50-52, 55)

$$p_k = \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

Odotusarvo on $\mathbb{E}(X) = np$ ja varianssi $D^2(X) = np(1-p)$.

2. Tarkastellaan satunnaismuuttujan $Z \sim N(0,1)$ jakaumaa.

(a) Hae mainituista taulukoista Z:n tiheysfunktion $\varphi(z)$ ja kertymäfunktion $\Phi(z)$ arvot, kun $z \in \{-3, -1, 0, 0.5, 2\}$.

(b) Hae Z:n fraktiilit z_p kun $p \in \{0.025, 0.1, 0.67, 0.95\}$.

3. Oletetaan, että suomalaisten naisopiskelijoiden populaatiossa kehon pituus X (ilman pyöristystä lähimpään kokonaiseen senttimetriin) noudattaa normaalijakaumaa odotusarvolla $\mu = 166$ cm ja varianssilla $\sigma^2 = 5^2$ cm^2.

(a) Laske seuraavat todennäköisyydet

(i) $\mathbb{P}(X \leq 150)$, (ii) $\mathbb{P}(150 < X \leq 180)$, (iii) $\mathbb{P}(X \geq 180)$.

(b) Laske X:n jakauman 95% **viitevälin** rajat eli 2.5% ja 97.5% fraktiilit, $x_{0.025}$ ja $x_{0.975}$, joiden väliin sijoittuu 95% jakaumasta.

4. Tarkastellaan toistokoetta, jossa heitetään arpanoppaa ja yksittäisessä heitossa kohdetapahtumana on $A =$ "silmäluku on 5 tai 6". Tällöin $p = \mathbb{P}(A) = 1/3$.

(a) Heitetään noppaa 6 kertaa. Olkoon $X =$ tapahtuman A esiintymiskertojen lukumäärä tässä heittosarjassa. Laske todennäköisyys sille, että $X \leq 1$.

(b) Heitetään noppaa 60 kertaa. Arvioi normaaliapproksimaatiolla (ks. Tuominen 1993, s. 121-123) todennäköisyyttä, että $X \leq 15$. Vertaa tarkkaan todennäköisyyteen, joka on $0.1071\ldots$.

Ota tämä tehtäväpaperi sekä ratkaisut mukaan myös M-harjoitukseen 1, jossa mm. toteutetaan tehtävien 2. – 4. vaatimat laskelmat käyttäen R:n työkaluja.

Kotitehtävien käsittelyn jälkeen toteutetaan datankeruu- ja mittausharjoitus ryhmän vetäjän johdolla.

Oulun yliopiston matemaattisten tieteiden laitos/tilastotiede

806113P TILASTOTIETEEN PERUSTEET, kl 2011 (Esa Läärä)

L-harjoitus 2, viikko 4 (to 27.1.): kotitehtävät

1. Kristiina Kuntun ja Teppo Huttusen *Korkeakouluopiskelijoiden terveystutkimus 2008* on pdf-muotoisena osoitteessa `http://www.yths.fi/filebank/587-45_OTT_Kunttu-Huttunen.pdf`. Tutkimuksessa käytetty kyselylomake löytyy tämän dokumentin sivuilta 351-370.

Tutustu kyselylomakkeen kysymyksiin n:o 8, 9, 12, 13, 15, 23, 51, 54, 55 ja niissä annettuihin vastausvaihtoehtoihin. Kerro kunkin em. kysymyksen kohdalla, mikä on vastaavan muuttujan (tai muuttujien, jos sama kysymys sisältää useampia muuttujia) mittaustaso t. mitta-asteikon tyyppi, kuin myös onko muuttuja diskreetti vai jatkuva.

2. Vertaillaan kahden auton A ja B nopeusmittarien toimivuutta. Kummallakin autolla ajettiin 6 kertaa tutkaan siten, että auton nopeusmittari osoitti joka kerran 100 km/h. Tarkasti kalibroidulla tutkalla saatiin tietoon auton todellinen nopeus. Seuraavassa on esitetty testin tulokset, jotka kertovat auton mittarilukeman ja todellisen nopeuden välisen erotuksen (km/h):

auto A:	4.4	5.0	4.6	4.5	4.8	4.6
auto B:	-1.0	-1.5	2.0	1.8	0.5	-1.8

(a) Kuvaa mittaustulokset sellaisella pistekuvioesityksellä, jonka avulla voidaan myös havainnollisesti vertailla nopeusmittarien luotettavuutta autojen välillä.

(b) Kommentoi autojen A ja B nopeusmittarien toimivuutta: kummalla näyttää olevan suurempi systemaattinen virhe eli harha ja kummalla suurempi satunnaisvirhe eli pienempi tarkkuus?

3. (Vanha tenttitehtävä.) Tilastotieteen perusteet A -kurssilla v 2010 toteutetussa datankeruu- ja mittausharjoituksessa 48 osallistujalta mitattuja diastolisen verenpaineen (eli "alapaineen") arvoja (mmHg) ja niiden jakaumaa toisella mittauskerralla (muuttuja `PAINEM2A`) kuvaa seuraava runko-lehtikuvio.

```
> stem(PAINEM2A)
  The decimal point is 1 digit(s) to the right of the |

   6 | 0188
   7 | 0111122233345789
   8 | 03344445566779
   9 | 00337888999
  10 | 2
  11 |
  12 | 1
  13 |
  14 | 1
```

Määrää seuraavien tunnuslukujen arvot, jotka kuvaavat diastolisen verenpaineen mittaustulosten empiiristä jakaumaa tässä populaatiossa:

(a) mediaani, minimi, maksimi, vaihteluväli,

(b) ala- ja yläkvartiili, kvartiiliväli ja kvartiilivälin pituus,

(c) aritmeettinen keskiarvo sekä keskihajonta, kun lisätietona annetaan, että mittaustulosten summa ja neliöpoikkeamien summa olivat

$$\sum_{i=1}^{n} x_i = 4034 \text{ mmHg}, \quad \sum_{i=1}^{n} (x_i - \bar{x})^2 = 10253.92 \text{ mmHg}^2.$$

Onko jakauma mielestäsi symmetrinen, oikealle vino vai vasemmalle vino?

4. Eräällä aiemmalla kurssilla toteutetussa datankeruu- ja mittausharjoituksessa yksi kysymys koski sitä, kuinka monta henkeä kaikkiaan oli kotitaloudessa, johon vastaaja itse kuului (muuttuja `KOTITAL`). Vastausten jakauma oli seuraavanlainen

kotitalouden koko	1	2	3	4	5	6	Yhteensä
vastaajia	21	9	7	7	4	2	50

(a) Havainnollista vastausten %-jakaumaa piirtämällä vastaava janakuvio (ks. tn-laskennan peruskurssi) eli piikkikuvio eli nuppineulakuvio.

(b) Määrää jakauman moodi, mediaani sekä ala- ja yläkvartiili.

(c) Laske kotitalouden koon aritmeettinen keskiarvo.

Oulun yliopiston matemaattisten tieteiden laitos/tilastotiede

806113P TILASTOTIETEEN PERUSTEET, kl 2011 (Esa Läärä)

L-harjoitus 2, viikko 4 (to 27.1.): tuntitehtävä

Oheisissa sanomalehtileikkeissä kerrotaan samasta äskettäin julkistetusta tutkimuksesta *Nuorten asuminen 2010 – Omaa kotia etsimässä*. Lue nämä uutistekstit huolellisesti ja yritä niiden perusteella löytää vastaukset seuraaviin kysymyksiin niin hyvin kuin mahdollista. Vertaa myös näiden kahden kirjoituksen sisältöä ja informatiivisuutta keskenään koskien sekä tutkimuksen ominaisuuksia että painotuksia tulosten esittelyssä.

(a) Mikä on tutkimuksen tavoite ja mitkä ovat sen pääkysymykset? Onko tutkimus kuvaileva vai syy-seuraussuhteita koskeva?

(b) Mikä on tutkimuksen kohdepopulaatio ja millaisista havaintoyksiköistä se koostuu? Kuinka suuri kohdepopulaatio on?

(c) Onko kyseessä kokonaistutkimus vai otantatutkimus? Jos se on otantatutkimus, niin millaista otantamenetelmää on käytetty ja kuinka suuri oli alkuperäinen otoskoko? Kuinka suuri oli vastauskato?

(d) Päättele tekstien pohjalta joitakin keskeisiä muuttujia, joita tutkimuksessa näyttää olevan mitattu ja analysoitu. Mitkä ovat näiden muuttujien mitta-asteikot?

(e) Mikä on tutkimuksen päätulos? Onko asioita, joita molemmat lehdet näyttävät yhteisesti painottavan, ja mitä eroja näet painotuksissa?

(f) Miten uutisointia mielestäsi voisi parantaa; mitä asioita tutkimuksen asetelmasta, menetelmistä ym. olisi toivottavaa kertoa tarkemmin?

Kaupunkimainen elämäntapa imaisee nuoren

Kaleva 15.1.2011

Tutkimuksen mukaan nuoret asuvat aiempaa ahtaammin

Piritta Rautavuori STT

HELSINKI Nuoret kaupunkilaiset haluavat säilyttää urbaanin elämäntapansa. Ympäristöministeriön perjantaina julkaiseman tutkimuksen mukaan kaupungeissa asuvista 18–29-vuotiaista suomalaisista entistä harvempi haaveilee maalle muutosta.

”Suurissa kaupungeissa asuvat haluavat entistä yksiselitteisemmin asua suurissa kaupungeissa”, summaa tutkija **Tiina Kupari** Nuorisoasuntoliitto ry:stä.

Kaupungissa asuminen tarkoittaa monelle nuorelle neliöistä tinkimistä samalla kun muu väestö elää yhä väljemmin. Asuintila on pienentynyt viidessä vuodessa etenkin Helsingissä.

”Jos nuoret haluavat asua entistä enemmän keskustoissa, joissa hinnat ovat korkeampia, neliöitä on väistämättä vähemmän”, totesi asuntoministeri **Jan Vapaavuori** (kok.).

> **”29-vuotiaista osa muistaa lamasta vielä jotakin, mutta 18-vuotiaille se on urbaanilegenda.”**
>
> **Jan Vapaavuori asuntoministeri (kok.)**

Hänen mukaansa Suomessa pitäisi ajatella, että asumisen laatu on muutakin kuin neliöitä. ”Jos aidosti arvostamme eheitä yhdyskuntarakenteita ja annamme arvoa ilmastokysymyksille, pitäisi keskustella siitä, että asumisen laatu lähtee muusta kuin neliöistä.”

Vapaavuori on pitkään puhunut eurooppalaisen kaupunkikulttuurin puolesta. Hän oli tutkimuksen tuloksista mielissään. ”Myös lapsiperheillä viisari värähtää urbaanin elämänmuodon suuntaan aiempaa enemmän. Se on erittäin positiivista.”

Nuoret suhtautuvat oman asunnon ostoon hyvin myönteisesti. Tutkimukseen vastanneista noin 70 prosenttia oli sitä mieltä, että asunnon ostaminen kannattaa aina. Yli 25-vuotiaista runsaat puolet kertoi asuvansa omistusasunnossa.

Korkojen alhaisuus on ruokkinut intoa ostaa oma asunto, mutta taustalla vaikuttanee myös se, ettei 1990-luvun alun lama kummittele nykyisten parikymppisten mielessä. ”29-vuotiaista osa muistaa lamasta vielä jotakin, mutta 18-vuotiaille se on urbaanilegenda, Vapaavuori sanoi.”

Noin puolet nuorista kertoi kiinnittävänsä huomiota rakennuksen energiatehokkuuteen valitessaan asuntoa. Hieman yli 40 prosenttia oli valmis maksamaan vihreästä sähköstä enemmän kuin tavallisesta.

”Ympäristöasioissa nuoret ovat todennäköisesti valveutuneempia kuin väestö keskimäärin”, Vapaavuori arvioi.

Suomalaisnuoret muuttavat kotoa varhain

Suomalaiset nuoret muuttavat pois kotoa keskimäärin 19-vuotiaana.
Vastaajista 60 prosenttia asui vuokralla ja 38 prosenttia omistusasunnossa.
Asuintilaa oli keskimäärin 32 neliötä henkeä kohden, kun vuonna 2005 neliöitä oli 33,3.
Nuoret ovat valmiita maksamaan asunnosta keskimäärin 180 000 euroa.
Asumismenot ovat keskimäärin 600 euroa kuussa. Menot ovat kasvaneet 19 prosentilla vuodesta 2005.
80 prosenttia oli sitä mieltä, että omat tulot riittävät hyvin asumiseen.
Lähde: Nuorten asuminen 2010 – Omaa kotia etsimässä -tutkimus

1. Miten ja missä asut nyt?
2. Miten suunnittelet asuvasi tulevaisuudessa?

HS 15.1.2011

Heikki Tani, 22, Helsinki
1. Asun vanhempieni omistamassa asunnossa Punavuoressa.
2. Lähitulevaisuudessa asun vuokra-asunnossa. Ehkä joskus myöhemmin hankin oman asunnon. Asuinpaikkani on todennäköisesti jokin muu kuin Helsinki. Viihdyn pienillä paikkakunnilla.

Sofia Rytky, 21, Helsinki
1. Asun yhdessä poikaystäväni kanssa Hoasin opiskelija-asunnossa Lassilassa.
2. Noin kymmenen vuoden kuluttua toivon asuvani omistusasunnossa. Olen kotoisin Oulun seudulta, mutta toiveasuinpaikkani on hiukan etelämpänä, esimerkiksi Jyväskylän seudulla.

Thomas Wahl, 26, Vantaa
1. Asun yhdessä tyttöystäväni kanssa omistusasunnossa Hiekkaharjussa.
2. Asumme todennäköisesti vielä aika pitkään nykyisessä asunnossa. Tulevaisuudessa asuntona on edelleen omistusasunto. Se on luultavasti lähempänä Helsingin keskustaa kuin nykyinen.

Kirsi Vuoristo, 20, Vantaa
1. Asun nyt vuokra-asunnossa Hiekkaharjussa yhdessä aviomieheni kanssa.
2. Tulevaisuudessa asumme omistusasunnossa. Olemme jo katselleet asuntoja ja ehkä hankimme sellaisen vuoden kuluessa. Se on todennäköisesti kerrostaloasunto Vantaalla.

Teemu Teräväinen, 26, Helsinki
1. Asun omistusasunnossa Etu-Töölössä. Hankin sen puolitoista vuotta sitten.
2. Myös tulevaisuudessa asun omistusasunnossa. Jos olen töissä Suomessa, asuinpaikkani on Helsinki. Asuinpaikka voi olla myös ulkomailla, jos sieltä löytyy sopivaa työtä.

Eine Siimes, 23, Helsinki
1. Nyt asumme kaupungin vuokra-asunnossa Tapanilassa.
2. Myöhemmin toiveena on oma asunto pääkaupunkiseudulta. Rivitalo olisi mieluisin.

TEKSTI: JORMA ERKKILÄ HS
KUVAT: SIRPA RÄIHÄ HS

Nuoret suosivat yhä enemmän omistusasumista

Kaupunkimainen elämä ja asuminen kiinnostavat 18–29-vuotiaita.

Marja Salmela HS

OMISTUSASUMISEN suosio kasvaa nuorten piirissä vuosi vuodelta. Nyt jo 70 prosenttia nuorista aikuisista pitää omaa asuntoa kannattavana hankintana.

Yli 25-vuotiaiden joukossa on jo enemmän omistusasukkaita kuin vuokralaisia, kertoo ympäristöministeriön julkaisema tutkimus "Omaa kotia etsimässä".

VUOKRALLA asuvistakin suurin osa pitää vaihetta väliaikaisena. Vain neljä prosenttia uskoo asuvansa vuokralla hyvin todennäköisesti pysyvästi.

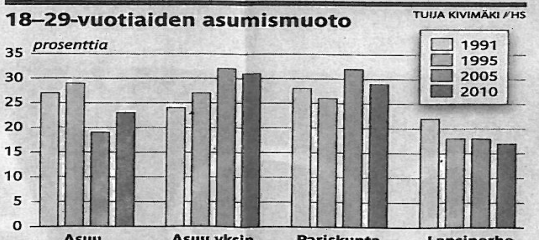
Tutkimusta varten haastateltiin yli tuhatta 18–29-vuotiasta eri puolilta maata viime vuonna.

Edellisen kerran vastaava tutkimus tehtiin vuosina 2005, 1995 ja 1991.

Runsas viidennes ilmoitti ostavansa asunnon vuoden sisällä ja noin 40 prosenttia kolmen vuoden aikana.

18–29-vuotiaiden asumismuoto

TUIJA KIVIMÄKI / HS



Suunnitelmissa on hankkia 2–4 huoneen asunto. Yksiöitä ei juuri haviteltu.

ASUNNOSTA nuoret olivat valmiita maksamaan keskimäärin 180 000 euroa ja pääkaupunkiseudulla jopa 200 000 euroa.

"Nykyinen, alhainen korkotaso varmasti vahvistaa omistusasumisen suosiota", arvelee asuntoministeri **Jan Vapaavuori** (kok).

Vapaavuori kehottaa nuoria pitämään lainanotossa pään kylmänä. "Moni elää maksukykynsä rajoilla."

Viime lama kutisti nuorten asumisväljyyttä, kun taas muu väestö asuu entistä väljemmin.

NYT NUORET asuvat ahtaammin kuin viisi vuotta sitten erityisesti Helsingissä ja muualla pääkaupunkiseudulla.

"Toisaalta yhä useampi nuori arvostaa neliöitä enemmän kaupunkimaista elämää lähellä kahviloita, kauppoja ja kulttuuripalveluja", kertoo tutkija **Tiina Kupari**.

Puolet asuisi mielellään monikulttuurisessa ympäristössä. Vain 15 prosenttia ei missään nimessä haluaisi naapureikseen maahanmuuttajia.

Ympäristöasenteita tiedusteltiin ensimmäistä kertaa.

Oulun yliopiston matemaattisten tieteiden laitos/tilastotiede

806113P TILASTOTIETEEN PERUSTEET A, kl 2010 (Esa Läärä)

L-harjoitus 3, viikko 5 (3.-5.2.): kotitehtävät

1. Jatkoa harjoituksen 2 tuntitehtävään. Katso uudelleen Helsingin Sanomien uutista tutkimuksesta *Nuorten asuminen 2010*. Siinä mm. kuvataan nuorten aikuisten asumismuodon jakaumaa ositetun pylväskuvion avulla. Luonnostele vaihtoehtoinen graafinen esitys alekkain asetettujen palkkikuvioiden avulla seuraten luennolla annetun esimerkin mallia. Vertaile omaa kuvaasi HS:n grafiikaan; kumpi on mielestäsi havainnollisempi?

2. Erään keskisuuren yrityksen työntekijöiden kuukausipalkat jakautuivat seuraavasti.

palkka (euroa)	1500-1999	2000-2499	2500-3499	3500-4999	5000-6999	Yhteensä
työntekijöitä	3	23	15	7	2	50

(a) Esitä kuukausipalkkojen jakauma graafisesti histogrammin avulla.

(b) Piirrä jakauman summakäyrä ja arvioi graafisesti mediaania sekä kvartiileja.

(c) Laske tämän aineiston pohjalta arviot seuraaville tunnusluvuille, jotka kuvaavat kuukausipalkkojen jakaumaa tässä yrityksessä. (i) aritmeettinen keskiarvo, (ii) keskihajonta ja (iii) geometrinen keskiarvo.

(d) Vertaa mediaania, aritmeettista keskiarvoa ja geometrista keskiarvoa keskenään; millaiseen suuruusjärjestykseen ne asettuvat?

(e) Onko keskihajonnalla konkreettista tulkintaa?

Huom. Rahasummat pyöristetään tavallisesti lähimpään alempaan euromäärään.

3. Piirrä laatikko-janakuvio seuraavista aineistoista ja lisäksi otoskertymäfunktion kuvaaja kohdan (b) aineistosta.

(a) Verenpaineen mittaustulokset ($n = 48$) L-harjoituksen 2 tehtävässä 3 (myös tämän kuvion piirtäminen sisältyi vanhaan tenttitehtävään).

(b) Kotitalouden koko L-harjoituksen 2 tehtävässä 4.

4. Erään sairaanhoitopiirin alueella vuoden 2009 aikana diagnosoitiin viidellä henkilöllä haimasyöpä. Näiden potilaiden elinajat diagnoosin jälkeen olivat pienimmästä suurimpaan 1, 2, 4, 7 ja 11 kk.

(a) Laske tämän potilasryhmän elinaikojen mediaani, kvartiilit, aritmeettinen keskiarvo ja keskihajonta.

(b) Jos pisimpään eläneen potilaan elinaika olisikin ollut 11 kk asemesta 22 kk, niin miten em. tunnusluvut muuttuisivat? Keskiarvoa ja keskihajontaa laskiessasi käytä hyväksesi tehtävässä 6. annettuja päivityskaavoja ja sitä, että 4 ensimmäisen havainnon keskiarvo ja varianssi ovat

$$\bar{x}_4 = \frac{1}{4}(1 + 2 + 4 + 7) = 3.5 \text{ kk}, \qquad s_4^2 = 7 \text{ kk}^2.$$

(c) Jos olisikin käynyt niin (kuten tämäntapaisissa lääketieteellisissä havaintoaineistoissa usein tapahtuu), että potilaiden seuruu päättyi ajankohtana, jolloin viides potilas oli 11 kk diagnoosin jälkeen vielä elossa, mutta kaikki muut potilaat olivat kuolleet, ja heidän elinaikansa olivat juuri nuo 1, 2, 4 ja 7 kk. Mitä tunnuslukuja voidaan tällaisessa tilanteessa laskea potilasryhmän elinaikojen jakaumalle? Onko keskiarvo laskettavissa ja miten?

Seuraavat tehtävät ovat ylimääräisiä harrastustehtäviä keskiarvon ja varianssin/keskihajonnan matemaattisista ominaisuuksista kiinnostuneille. Kaikkien kuitenkin kannattaa käydä läpi annetut lopputulokset

5. Olkoon $x_1, x_2 \ldots, x_n$ muuttujan x havaittuja arvoja n havaintoyksiköllä. Olkoon muuttuja y muuttujan x lineaarinen muunnos; ts. $y_i = a+bx_i,\ i = 1, 2, \ldots, n,$, jossa a ja b ovat reaalilukuja. Merkitään kummankin muuttujan keskiarvoa ja keskihajontaa tavanomaiseen tapaan

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

ja vastaavasti määritellään $\bar{y}$ sekä s_y. Olkoot myös $z_i = (x_i - \bar{x})/s_x$, $i = 1, 2, \ldots, n$; ts. z_i:t ovat muuttujan x **standardoidut** tai **normitetut** arvot. Osoita, että

(a) $\bar{y} = a + b\bar{x}$, (b) $s_y = |b|s_x$, (c) $\bar{z} = 0$ ja $s_z = 1$.

6. Oletetaan, että n havaintoa sisältävästä havaintovektorista $X_n = (x_1, \ldots, x_n)$ on laskettu aritmeettinen keskiarvo $\bar{x}_n$ ja varianssi s_n^2, joista jälkimmäiselle on alla annettu sekä määritelmän mukainen kaava että oikeanpuoleisen lausekkeen mukainen "laskukonekaava":

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right].$$

Tämän jälkeen saadaan lisähavainto x_{n+1}. Osoita, että lisähavainnolla täydennetyn vektorin $X_{n+1} = (x_1, \ldots, x_n, x_{n+1})$ alkioiden

(a) aritmeettinen keskiarvo $\bar{x}_{n+1}$ saadaan vanhasta keskiarvosta $\bar{x}_n$ ja uudesta havainnosta x_{n+1} seuraavalla päivityskaavalla

$$\bar{x}_{n+1} = \frac{n}{n+1}\bar{x}_n \ + \ \frac{1}{n+1}x_{n+1} = \bar{x}_n \ + \ \frac{1}{n+1}(x_{n+1} - \bar{x}_n)$$

(b) varianssille s_{n+1}^2 pätee päivityskaava

$$s_{n+1}^2 = \left(\frac{n-1}{n}\right)s_n^2 \ + \ \frac{1}{n+1}(x_{n+1} - \bar{x}_n)^2.$$

Vihje. Tämän kaavan pätevyyden todistamisessa kannattaa lähteä liikkeelle "laskukonekaavasta" alkamalla purkaa sen sisältämät neliösumma- ja summalausekkeet vektorista X_n ja havainnosta x_{n+1} riippuviin komponentteihin, ja lopuksi voi soveltaa kohdan (a) tulosta.

Huom. Kunnollisissa tilastollisissa ohjelmistoissa keskiarvon ja varianssin laskenta-algoritmit perustuvat tämäntyyppisiin päivityskaavoihin eikä suinkaan havaintojen summaan ja neliösummaan nojautuviin "laskukonekaavoihin", jotka isoilla luvuilla ja havaintomäärillä voivat olla hyvin epästabiileja. Testaapa esim. vektorilla $X_3 = (9000000001, 9000000002, 9000000003)$, jonka alkioiden varianssi ja keskihajonta ovat molemmat 1.

Oulun yliopiston matemaattisten tieteiden laitos/tilastotiede

806113P TILASTOTIETEEN PERUSTEET, kl 2011 (Esa Läärä)

L-harjoitus 4, viikko 6 (10.2.): kotitehtävät

Oheisessa liitteessä on Kalevan lehtiuutinen sekä tiivistelmä ja tulostaulukoita eräästä tammikuussa julkaistusta tutkimuksesta aiheena C-vitamiinin yhteys ylähengitystieinfektio- (URI) eli flunssaepisodien ilmaantuvuuteen, kestoon ja oireiden vaikeusasteeseen. Lue liitteen kaikki osat huolellisesti ja yritä vastata kysymyksiin 1. ja 2. niin hyvin kuin annetun informaation pohjalta kykenet.

1. Tutkimuksen kysymyksenasettelu, populaatio, asetelma ja menetelmät:

(a) Miten muotoilisit tutkimuksen pääkysymykset täsmällisesti?

(b) Mikä on tutkimuksen kohdepopulaatio; millaisista havaintoyksiköistä se koostuu? Voidaanko tutkimuksessa mukana olleita henkilöitä pitää satunnaisotoksena kohdepopulaatiosta?

(c) Miten luonnehdit tutkimusasetelmaa: onko se kokeellinen vai epäkokeellinen? Mikä rooli on satunnaistuksella (*randomization*), kaksoisnaamioinnilla eli "-sokkoutuksella" (*double-blind*) ja lume-kontrollilla (*placebo-control*)? Näyttävätkö vertailtavat ryhmät olleen perusominaisuuksiltaan vertailukelpoisia keskenään?

(d) Mitkä olivat tutkimuksen keskeiset tulos- eli vastemuuttujat ja millä mitta-asteikolla niitä mitattiin? Mitkä olivät tärkeimmät selittävät tekijät?

2. Tutkimuksen tulokset ja niiden tulkinta:

(a) Mitä tunnuslukuja tutkimuksen päätulosten esittelyssä (Table 3) käytettiin? Mitä voit päätellä tutkittavien muuttujien jakauman symmetrisyydestä tai vinoudesta kussakin ryhmässä raportoitujen tunnuslukujen suuruuksien perusteella? Mitä muita tunnuslukuja olisit toivonut raportoitavan muuttujien jakaumasta?

(b) Mikä tai mitkä oli(vat) tutkimuksen päätulo(k)s(et)? Ovatko Kalevan uutisen sekä itse tutkimusraportin tiivistelmän tekstit sopusoinnussa Table 3:n kanssa? Kuinka suuria olivat päätuloksiin liittyvät virhemarginaalit?

(c) Miten tulkitset tuloksia? Osoittavatko ne vakuuttavasti, että C-vitamiini ei vähennä uusien episodien ilmaantuvuutta? Onko saatu näyttö vakuuttava sen puolesta, että C-vitamiinilla olisi vaikuttavuutta vain pojilla mutta ei tytöillä? Kuinka laajaan populaatioon tulokset voisivat olla yleistettävissä?

Tarkastellaan seuraavaksi puolueiden kannatusosuuksia ja niiden arviointia viime aikoina lehdissä ilmestyneiden uutisten valossa. Lähdemateriaalina ovat (i) tämän viikon luennoilla jaettu materiaali, joka sisältää kuvauksen Taloustutkimuksen Ylelle tekemistä puolukannatusmittauksista, niiden menetelmistä ja tuoreimmista tuloksista, kuin myös Helsingin Sanomien uutisen 26.1. TNS Gallupilla teettämästään puoluekannatusmittauksen tuloksista, ja (ii) luvun 3 luentomonisteen toisen osan lopussa olevan lehtileikkeen Kalevassa 14.1. ilmestyneestä uutisesta, jonka pääaiheena oli Vanhasen esteellisyys RaY:n valtionavuista päätettäessä, mutta jossa raportoitiin myös vastaajien puoluekannatuksen jakauma.

Satunnaisotantaan perustuvassa kannatusmittauksessa yksittäisen puolueen kannatusosuuden tavanomainen virhemarginaali eli 95% **luottamusväli** lasketaan seuraavalla periaatteella. Merkitään

n = kaikkien niiden otokseen poimittujen ja tutkimukseen osallistuneiden henkilöiden lukumäärä, jotka ilmaisivat kannattavansa jotain puoluetta,

m_k = puoluetta k kannattaneiden havaittu lukumäärä em. n henkilön joukossa,

p_k = m_k/n = puolueen k otoksesta arvioitu kannatusosuus.

Näiden pohjalta saadaan arvioidun kannatusosuuden **keskivirhe** $\mathrm{SE}(p_k)$, ja lopuksi likimääräisen 95% luottamusvälin (CI) ala- ja yläraja todelliselle kannatusosuudelle

$$\mathrm{SE}(p_k) = \sqrt{\frac{p_k(1-p_k)}{n}}, \qquad \mathrm{CI} = [p_k - 1.96 \times \mathrm{SE}(p_k),\ p_k + 1.96 \times \mathrm{SE}(p_k)],$$

jossa 1.96 on $N(0,1)$-jakauman 97.5% fraktiili.

Vastaa annettujen taustatietojen pohjalta kysymyksiin 3. ja 4. niin hyvin kuin voit.

3. Kalevan uutinen 14.1. kansalaisten mielipiteistä koskien Vanhasen toimien lainmukaisuutta.

(a) Laske tässä tutkimuksessa havaittu Perussuomalaisten kannatusosuus ja sen 95% luottamusväli kaikkien puoluekannatuksensa ilmaisseiden vastaajien joukossa. Vertaa Taloustutkimuksen ja TNS Gallupin antamiin tuoreisiin kannatusosuuksiin ja niiden virhemarginaaleihin. Mitä havaitset ja miten tulkitset? Onko Kalevan uutisoima tulos mielestäsi uskottava?

(b) Laske samasta aineistosta vastaavalla tavalla kannatusosuus ja sen 95% luottamusväli myös Suomen Keskustalle. Vertaa virhemarginaalin leveyttä kohdan (a) virhemarginaaliin. Vertaa myös tätä tulosta Taloustutkimuksen ja TNS Gallupin tuloksiin Keskustan kannatuksesta. Ovatko eri tutkimusten tulokset sopusoinnussa vai keskenään ristiriidassa?

4. Taloustutkimuksen ja TNS Gallupin puoluekannatusarviot.

(a) Taloustutkimus kertoo, kuinka moni sen tammikuussa haastattelemista ihmisistä ilmaisi kannattavansa jotain puoluetta. Helsingin Sanomien uutinen ei anna vastaavaa lukua TNS Gallupin tammikuisen otoksen osalta. Minkä tietojen perusteella voidaan kuitenkin uskottavasti arvioida, että TNS Gallupin tammikuun mittauksessa jotakin puoluetta kannattaneiden lukumäärä oli suuruusluokaltaan pyöreästi n. 1600?

(b) Lehtiuutisissa raportoidut virhemarginaalit koskevat sellaisenaan vain suurimpien puolueiden kannatusosuuksia, jotka ovat kokoluokkaa 20%. Kuinka suuri virhemarginaali on TNS Gallupin otoksen pohjalta laskettuna sellaisilla puolueilla (kuten RKP ja Kristillisdemokraatit), joiden kannatusosuus on n. 4%?

(c) Lue huolellisesti Taloustutkimuksen kannatusarvioiden menetelmäkuvauksessa otsikon "Kannatusarvion laskentatapa" alla oleva teksti. Mitä sen perusteella päättelet Perussuomalaisten kannattajaksi ilmoittautuneiden lukumäärästä m_{PS} ja osuudesta p_{PS} tammikuun kyselyyn vastanneiden $n = 2005$ henkilön keskuudessa; onko siinä havaittu osuus p_{PS} ollut täsmälleen 16.6%, enemmän kuin 16.6% vai vähemmän kuin 16.6%? Perustele.

Kaleva 15.1.2011

C-vitamiini lyhensi uimarien flunssia

C-vitamiini lyhensi flunssien kestoa nuorilla israelilaisilla kilpauimareilla, ilmenee European Journal of Pediatrics -lehdessä julkaistusta israelilais-suomalaisesta tutkimuksesta.

C-vitamiini lyhensi flunssien kestoa kuitenkin vain pojilla, joilla flunssat lyhenivät noin 40 prosenttia. Tyttöjen flunssia C-vitamiini ei lyhentänyt.

C-vitamiini ei myöskään vähentänyt flunssien lukumäärää tytöillä eikä pojilla.

Tutkimukseen osallistui 42 kilpauimaria, iältään 12-17-vuotiaita; poikia oli 24 ja tyttöjä 18. Kaikki osallistujat käyttivät vähintään 15 tuntia viikossa uintiharjoituksiin.

C-vitamiinin vaikutusta flunssaan on tutkittu paljon.

Aikaisemmissa tutkimuksissa on havaittu C-vitamiinin lyhentävän flunssan kestoa ja estävän flunssia poikkeuksellisen rasitetuilla ihmisillä, kuten maratonjuoksijoilla.

The effect of vitamin C on upper respiratory infections in adolescent swimmers: a randomized trial

Naama W. Constantini, Gal Dubnov-Raz, Ben-Bassat Eyal, Elliot M. Berry, Avner H. Cohen, and Harri Hemilä

European Journal of Pediatrics 2011; 170: 59–63.

Abstract

The risk of upper respiratory infections (URIs) is increased in people who are under heavy physical stress, including recreational and competitive swimmers. Additional treatment options are needed, especially in the younger age group. The aim of this study was to determine whether 1 g/day vitamin C supplementation affects the rate, length, or severity of URIs in adolescent swimmers. We carried out a randomized, double-blind, placebo-controlled trial during three winter months, among 39 competitive young swimmers (mean age 13.8 ± 1.6 years) in Jerusalem, Israel. Vitamin C had no effect on the incidence of URIs (rate ratio = 1.01; 95% confidence interval (CI) = 0.70–1.46). The duration of respiratory infections was 22% shorter in vitamin C group, but the difference was not statistically significant. However, we found a significant interaction between vitamin C effect and sex, so that vitamin C shortened the duration of infections in male swimmers by 47% (95% CI: −80% to −14%), but had no effect on female swimmers (difference in duration: +17%; 95% CI: −38% to +71%). The effect of vitamin C on the severity of URIs was also different between male and female swimmers, so that vitamin C was beneficial for males, but not for females. Our study indicates that vitamin C does not affect the rate of respiratory infections in competitive swimmers. Nevertheless, we found that vitamin C decreased the duration and severity of respiratory infections in male swimmers, but not in females. This finding warrants further research.

Table 1 Clinical and training data of athletes in placebo and vitamin C groups

	Placebo	Vitamin C
Number of participants (*n*)	18	21
Age, years	13.9±1.8	13.7±1.5
Males (*n*)	10	12
Swimming duration, h/week	11±4	11±4
Swimming distance, km/week	34±13	30±12
Fitness training, h/week	2.3±1.9	1.6±1.8
Other training[a], h/week	2.3±2.0	1.7±1.6

Values are *n* or mean±SD, as appropriate

[a] e.g., Ball games, cycling, school activities

Table 2 The number of upper respiratory infections by treatment group

	Placebo	Vitamin C	RR (95% CI)
Number of participants	18	21	
Number of URI episodes	54	64	1.01 (0.70, 1.46)

RR rate ratio, *URI* upper respiratory infection

Table 3 The duration and severity of upper respiratory infections by treatment group

	Placebo	Vitamin C	Difference (95% CI)	Test of interaction[b] *P*
Number of URI episodes[a]	43	55		
Duration of URI episodes (days) (mean±SD)	8.9±7.8	6.9±5.4	−2.0 (−4.6, +0.7)	
Males (21+30 episodes)	10.4±7.1	5.5±5.0	−4.9 (−8.4,−1.5)	0.003
Females (22+25 episodes)	7.4±8.2	8.6±5.5	+1.2 (−2.8, +5.3)	
Severity of URI episodes (Mean±SD)	59±87	43±45	−16 (−43, +11)	
Males (21+30 episodes)	66±85	26±30	−40 (−75,−6)	0.003
Females (22+25 episodes)	52±89	64±51	+12 (−30, +54)	

URI upper respiratory infection

[a] Data for this table is restricted to the first four URI episodes of the participants, see "Methods" section

[b] The test of interaction was carried out with log transformed duration and severity

Oulun yliopiston matemaattisten tieteiden laitos/tilastotiede

806113P TILASTOTIETEEN PERUSTEET, kl 2011 (Esa Läärä)

L-harjoitus 5, viikko 7 (to 17.2.): kotitehtävät

1. Aku Ankka ja Hannu Hanhi pelaavat noppapeliä. Aku alkaa epäillä, että Hannun noppa on painotettu, koskapa Hannun 30 ensimmäisen heiton joukossa ainoastaan 2 kertaa silmäluvuksi tuli 'yksi', kun taas silmäluku 'kuusi' esiintyi paljon useammin. Merkitään $\theta = \mathbb{P}(A)$, jossa $A =$ "yksittäisessä heitossa silmäluvuksi saadaan 'yksi' ".

(a) Laske parametrin θ suurimman uskottavuuden estimaatti.

(b) Jos noppa on reilu, niin voidaan odottaa, että $\theta = \mathbb{P}(A) = 1/6$. Pidetään tätä nollahypoteesina. Laske saamastasi aineistosta tämän nollahypoteesin testaamisessa käytettävän testisuureen Z arvo ja arvioi normaalijakauman taulukkoa hyväksi käyttäen vastaava P-arvo. Mitä päättelet? Antaako havaittu tulos näyttöä nollahypoteesia vastaan? Saako nollahypoteesi tukea?

(c) Laske 95% likimääräinen luottamusväli parametrille θ käyttäen yksinkertaisinta laskukaavaa, joka perustuu muokkaamattomaan su-estimaattiin ja keskivirheeseen. Mitä havaitset? Onko luottamusväli looginen?

(d) Laske 95% likimääräinen luottamusväli parametrille θ soveltaen nyt AC-menetelmää, joka perustuu muokattuihin tunnuslukuihin $\widetilde{\theta}$ ja $\mathrm{SE}(\widetilde{\theta})$. Onko luottamusväli nyt looginen? Kuinka leveä virhemarginaali on?

2. Taloustutkimuksen tammikuussa 2011 julkistamassa puoluekannatusmittauksessa raportoitiin SDP:n kannatusosuudeksi 18.9% niiden 2000 henkilön joukossa, jotka ilmoittivat kannattavansa jotakin puoluetta. Vuoden 2007 eduskuntavaaleissa SDP sai kaikista äänistä 21.4%.

(a) Laske SDP:n kannatusosuuden 95% luottamusväli 1/2011 suoritetun kannatusmittauksen tulosten pohjalta.

(b) Onko näyttöä siitä, että SDP:n kannatusosuus olisi muuttunut siitä, mikä se oli vuoden 2007 eduskuntavaaleissa? Testaa tätä nollahypoteesia, arvioi vastaava P-arvo ja tulkitse tulokset.

3. Puoluekannatusmittauksen virhemarginaalin leveys (95% luottamusvälin ylä- ja alarajan erotus) sellaisen puolueen kohdalla, jonka kannatusosuus on n. 20 %, on Taloustutkimuksen otoksessa ($n \approx 2000$ puoluekantansa ilmaissutta) n. 2×1.7 %-yksikköä, ja TNS Gallupin otoksessa ($n \approx 1600$) se on n. 2×2 %-yksikköä. Kuinka suuri otos tarvittaisiin kaikkiaan (mukaan lukien myös ne, jotka eivät kerro kannattavansa mitään puoluetta ja joita tyypillisesti on n. kolmasosa haastatelluista) tällaisessa kannatusmittauksessa, jos halutaan, että n. 20% kannatuksella virhemarginaalin kokonaisleveys olisi vain

(a) 2 %-yksikköä, (b) 1 %-yksikkö?

4. Luvun 4 luentomonisteen s. 10 keskellä kerrotaan, kuinka $100(1-\alpha)$ % luottamusvälin ala- ja ylärajat mallin $\mathrm{Bin}(n, \theta)$ parametrille θ voidaan laskea ns. Wilsonin menetelmällä ratkaisemalla θ':n suhteen 2. asteen yhtälö

$$\frac{(\widehat{\theta} - \theta')^2}{\theta'(1-\theta')/n} = z^2_{1-\alpha/2},$$

jossa z_u on $N(0,1)$-jakauman u-fraktiili, $0 < u < 1$. (Kun esimerkiksi $\alpha = 0.05$, tämä fraktiili on $z_{0.975} = 1.96$).

Johda Wilsonin menetelmää noudattavan luottamusvälin ala- ja ylärajojen lausekkeet ratkaisemalla tämä yhtälö.

5. Kertaustehtävä todennäköisyyslaskennan peruskurssin asioista: Oletetaan, että naispuolisten korkeakouluopiskelijoiden perusjoukossa pituus noudattaa normaalijakaumaa odotusarvolla $\mu = 166.5$ cm ja varianssilla $\sigma^2 = 5^2$ cm^2. Poimitaan tästä joukosta n henkilön satunnaisotos ja merkitään $X_i = i$:nnen otosyksilön pituus. Satunnaisotannan perusteella voidaan olettaa, että kullakin $i = 1, \ldots, n$ on $X_i \sim N(166.5, 5^2)$ toisista riippumatta. Merkitään n:n havaintoon perustuvaa otoskeskiarvoa $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

(a) Kuinka suuri on otoskeskiarvon $\bar{X}_n$ (otanta)jakauman odotusarvo $\mathbb{E}(\bar{X}_n)$?

(b) Jos otoskoko on $n = 25$, niin kuinka suuria ovat $\bar{X}_n$:n jakauman varianssi $\mathrm{var}(\bar{X}_n)$ ja keskihajonta $\mathrm{SD}(\bar{X}_n) = \sqrt{\mathrm{var}(\bar{X}_n)}$.

(c) Jos edelleen $n = 25$, niin kuinka suuria ovat $\bar{X}_n$:n jakauman fraktiilit $\xi_{0.025}$ ja $\xi_{0.975}$, jossa u-fraktiilille ξ_u pätee: $\mathbb{P}(\bar{X}_n \leq \xi_u) = u$, kun $u \in]0,1[$.

(d) Kuinka suuri pitää otoskoon n vähintään olla, jotta otantajakauman teoreettinen keskihajonta $\mathrm{SD}(\bar{X}_n)$ olisi korkeintaan 0.5 cm?

Oulun yliopiston matemaattisten tieteiden laitos/tilastotiede

806113P TILASTOTIETEEN PERUSTEET, kl 2011 (Esa Läärä)

L-harjoitus 6, viikko 8 (24.2.): kotitehtävät

1. Pari viikkoa sitten toteutetuissa nastanheittotalkoissa eri harjoitusryhmissä saatiin seuraavat yhdistetyt tulokset.

Ryhmä	Heittäjiä	Heittoja	Jäi selälleen
1 (Hanna)	7	175	84
2 (Hanna)	9	225	131
3 (Päivi)	15	375	219
Yhteensä	31	775	434

Merkitään θ_k = "todennäköisyys, että nasta jää selälleen yksittäisessä heitossa harjoitusryhmässä k", jossa $k = 1, 2, 3$.

(a) Laske θ_k:n piste-estimaatit, keskivirheet ja likimääräiset 95% luottamusvälit (yksinkertaisella kaavalla) erikseen kahdessa ensimmäisessä ryhmässä; ts. $k = 1, 2$

(b) Verrataan ryhmien 1 ja 2 tuloksia keskenään vertailuparametrin $\delta = \theta_1 - \theta_2$ avulla. Laske δ:n piste-estimaatti, keskivirhe ja likimääräinen 95% luottamusväli (yksinkertaisella kaavalla).

(c) Nollahypoteesina voidaan pitää $H_0 : \theta_1 = \theta_2 (= \theta_3)$, eli erityisesti ryhmien 1 ja 2 vertailussa $H_0 : \delta = 0$. Testaa tätä nollahypoteesia laskemalla testisuureen arvo (yksinkertaisella kaavalla) ja likimääräinen P-arvo. Mitä päättelet: onko tulos sopusoinnussa H_0:n kanssa, vai antavatko havainnot näyttöä siitä, että ryhmässä 1 heittotapa olisi ollut sen verran erilainen kuin ryhmässä 2, että selälleen jäämisen todennäköisyydet olisivat olleet erisuuruiset?

2. Jatkoa L-harjoituksen 5 tehtävään 2. Taloustutkimuksen puoluekannatusmittauksessa 1/2011 raportoitiin siis SDP:n kannatusosuudeksi 18.9% niiden 2000 henkilön joukossa, jotka ilmoittivat kannattavansa jotakin puoluetta. Maaliskuussa 2007 toteutetussa vastaavassa tutkimuksessa ja yhtä suuressa otoksessa SDP:n havaittu kannatusosuus oli 21.4%.

(a) Laske SDP:n kannatusosuuden 95% luottamusväli maaliskuussa 2007.

(b) Laske piste-estimaatti ja 95% luottamusväli SDP:n kannatusosuuksien erotukselle maaliskuun 2007 ja tammikuun 2011 välillä. Vertaa erotuksen virhemarginaalin leveyttä yksittäisten osuuksien virhemarginaaleihin.

(c) Testaa nollahypoteesia, jonka mukaan SDP:n todellinen kannatusosuus olisi pysynyt täsmälleen samana tammikuussa 2011 kuin mitä se oli maaliskuussa 2007: laske testisuureen arvo, arvioi vastaava P-arvo ja tulkitse tulokset. Vertaa siihen tulokseen, jonka sait edellisen harjoituksen tehtävästä 2.(b).

3. Suomen naispuolisten korkeakouluopiskelijoiden perusjoukossa systolinen verenpaine (mittayksikkönä elohopeamillimetri eli mmHg) noudattaa jotain jatkuvaa jakaumaa odotusarvolla μ ja varianssilla σ^2, jotka ovat tuntemattomia. Tämän kurssin sekä kl 2010 pidetyn vastaavan kurssi alkuvaiheessa toteutetussa datankeruu- ja mittausharjoituksessa yhteensä 39 naispuolisen osallistujan ensimmäisissä verenpainemittauksissa saatu systolisen verenpaineen keskiarvo oli 120 mmHg ja keskihajonta 19 mmHg. Havaintojen runko-lehtikuvio oli seuraavanlainen.

```
> stem(PAINEM1Y[sukup=="nainen"])
  The decimal point is 1 digit(s) to the right of the |

   9 | 1234689
  10 | 478889
  11 | 003668
  12 | 011126679
  13 | 0669
  14 | 14688
  15 | 0
  16 | 8
```

Jos voidaan olettaa, että tämä joukko on tarpeeksi edustava otos kohdepopulaatiosta, niin vastaa seuraaviin kysymyksiin:

(a) Mitkä ovat odotusarvon μ ja varianssin σ^2 piste-estimaatit?

(b) Kuinka suuri on tässä aineistossa systolisen verenpaineen mittaustulosten keskiarvon keskivirhe? Laske sen pohjalta odotusarvon μ luottamusvälin ala- ja ylärajat kahdella eri luottamustasolla: 90 % sekä 95%?

(c) Edellisessä kohdassa laskettu luottamusväli on periaatteessa "tarkka", eli sen otantajakauman teorianmukaiset ominaisuudet pätevät, jos kohdemuuttujaa koskevien havaintojen voi olettaa noudattavan tavanomaista mallioletusta. Mikä oli tämä mallioletus? Mitä mahdollisia ongelmia sen sopivuudessa voi olla tähän tilanteeseen? Miten keskeinen raja-arvolause (ks. todennäköisyyslaskennan peruskurssi) vaikuttaa luottamusvälin pätevyyteen, vaikka mallioletus ei sellaisenaan pitäisi paikkaansa?

4. Keski-ikäisten miesten ryhmälle ($n = 16$) tehtiin rasitustesti. Miesten verenpaineet mitattiin sekä ennen rasitusta että rasituksen jälkeen. Systolisen verenpaineen (mmHg) mittaustulokset olivat henkilöittäin seuraavat:

Henkilö	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Ennen	148	142	136	134	138	140	132	144	128	170	162	150	138	154	126	116
Jälkeen	152	152	134	148	144	136	144	150	146	174	162	162	146	156	132	126
Erotus	+4	+10	−2	+14	+6	−4	+12	+6	+18	+4	0	+12	+8	+2	+6	+10

Mittaustulosten erotusten "jälkeen – ennen" keskiarvo oli +6.6 mmHg ja keskihajonta 6.0 mmHg.

Oletetaan, että tällaisessa asetelmassa rasituksen aiheuttamat systolisen verenpaineen muutokset noudattavat ao. kohdepopulaatiossa jakaumaa, jonka odotusarvo on Δ ja varianssi τ^2 ovat tuntemattomat.

(a) Laske 95 % luottamusväli parametrille Δ.

(b) Tarkastellaan nollahypoteesia $H_0 : \Delta = 0$. Laske havainnoista vastaavan testisuureen arvo sekä P-arvo.

(c) Miten tulkitset tuloksia? Onko näin asetettu nollahypoteesi testaamisen väärti?

5. Lukion pitkän matematiikan oppikirjassa on seuraava esimerkki:

"Erään lukion 2. vuosikurssin matematiikan ryhmän [nais]opiskelijoiden kengännumeroista on frekvenssitaulukko sivulla 15 [ks. alla]. Arvioi tämän perusteella, mikä on lukion 2. vuosikurssin naisopiskelijoiden kengännumeroiden keskiarvo ja keskihajonta.

Naisten kengännumero	Frekvenssi
37	3
38	7
39	10
40	3
41	1

Ratkaisu: Koska kyseessä on otos, lasketaan otoskeskiarvo ja otoskeskihajonta. Otoskeskiarvo on

$$\bar{x} = \frac{3 \cdot 37 + 7 \cdot 38 + 10 \cdot 39 + 3 \cdot 40 + 1 \cdot 41}{24} = 38.666\ldots$$

Otosvarianssi on

$$s^2 = \frac{3 \cdot (37 - 38.67)^2 + 7 \cdot (38 - 38.67)^2 + \cdots + 1 \cdot (41 - 38.67)^2}{24 - 1} = 1.014\ldots$$

Otoskeskihajonta on $s = \sqrt{s^2} = \sqrt{1.014\ldots} = 1.007\ldots$

Siis Suomen lukioiden 2. vuosikurssin naisopiskelijoiden kengännumeroiden keskiarvo on $\mu \approx 39$ ja keskihajonta $\sigma \approx 1$."

(a) Mitkä osat kirjan esittämässä ratkaisussa ovat oikein ja järkeviä?

(b) Mitkä ratkaisun yksityiskohdat ovat pielessä?

(c) Jos voidaan olettaa, että taulukon havainnot muodostavat edustavan otoksen Suomen naispuolisten 2. vuosikurssin lukiolaisten kengännumeroista, niin laske likimääräinen 95% luottamusväli kengännumeron jakauman odotusarvolle μ tässä populaatiossa soveltaen luennoilla opetettua menetelmää (ks. luentomonisteen "Luku 5 ..." alaluku 5.4 esimerkkeineen).

(d) Mitä mahdollisia ongelmia voit havaita mallioletuksen, johon luottamusvälin nimelliset ominaisuudet nojaavat, sopivuudessa kengännumeroiden jakaumaa kuvaamaan? Miten arvioit keskeisen raja-arvolauseen vaikuttavan luottamusvälin pätevyyteen?

Oulun yliopiston matemaattisten tieteiden laitos/tilastotiede

806113P TILASTOTIETEEN PERUSTEET, kl 2011 (Esa Läärä)

L-harjoitus 7, viikko 9 (3.3.): kotitehtävät

HUOM. Tiistaina 8.3. klo 12.30 alkaen salissa L6 pidetään kurssin ylimääräinen kertaus- ja kyselyluento, joka saattaa olla hyödyllinen seuraavana maanantaina 14.3. klo 14-18 pidettävään loppukuulusteluun valmistautumisen kannalta.

1. Data-analyysin perusmenetelmien kurssilla sl 2009 toteutettiin pienimuotoinen koe, jossa sovellettiin täydellisesti satunnaistettua rinnakkaisten ryhmien asetelmaa. Kokeen tehtävänä oli selvittää lyhytaikaisen fyysisen ponnistelun vaikutusta sydämen lyöntitiheyteen eli syketaajuuteen (mittayksikkönä lyöntiä per minuutti).

Kurssin osallistujat satunnaistettiin kahteen ryhmään. Aluksi kumpikin ryhmä istui rauhallisesti paikoillaan muutaman minuutin, jotta saavutettaisiin tavanomainen leposykkeen taso, joka sitten mitattiin ja kirjattiin muuttujaan `alkusyke`. Seuraavaksi koeryhmän yksilöt ($n = 9$) suorittivat viisi (5) kertaa "istualtaan ylös ja takaisin" -liikettä, kun taas vertailuryhmän jäsenet ($n = 8$) istuivat paikoillaan. Välittömästi tämän jälkeen kukin koehenkilö mittasi ja kirjasi oman syketaajuutensa uudelleen muuttujaan `loppusyke`. Tulokset olivat seuraavat:

	ryhma	sukup	alkusyke	loppusyke
1	koe	nainen	80	90
3	koe	nainen	66	72
8	koe	nainen	74	80
14	koe	nainen	86	94
6	koe	mies	58	68
10	koe	mies	80	86
11	koe	mies	70	80
16	koe	mies	74	90
19	koe	mies	72	80
13	vertailu	nainen	76	74
15	vertailu	nainen	80	76
17	vertailu	nainen	76	72
2	vertailu	mies	72	68
4	vertailu	mies	90	92
5	vertailu	mies	90	94
9	vertailu	mies	78	76
12	vertailu	mies	76	72

Seuraavassa joitakin tunnuslukuja sekä alkusykkeen että loppusykkeen (per min) jakaumista koe- ja vertailuryhmissä:

```
> with(syk, round(tapply(alkusyke, ryhma, mean), 2))

     koe vertailu
   73.33    79.75

> with(syk, round(tapply(alkusyke, ryhma, sd), 2))
```

```
     koe vertailu
    8.31     6.71

> with(syk, round(tapply(loppusyke, ryhma, mean), 2))

     koe vertailu
   82.22    78.00

> with(syk, round(tapply(loppusyke, ryhma, sd), 2))

     koe vertailu
    8.63     9.62
```

Tutkitaan loppusykkeen jakauman eroa koeryhmän ja vertailuryhmän välillä.

(a) Muotoile malli, jolla analysoidaan loppusykkeen keskimääräistä eroa koekäsittelyn ja vertailukäsittelyn välillä: jakaumaoletus ja parametrit. Minkä suuntaiseksi odotusarvojen erotusta voisi ennustaa ilmiötä koskevan fysiologisen tiedon valossa?

(b) Piirrä loppusykkeen mittaustulokset ryhmittäin alekkaisiin pistekuvioihin. Mitä havaintoja teet jakaumien sijainnista, hajonnasta ja mahdollisesta vinoudesta? Miten arvioit malliolettusten realistisuutta tässä aineistossa? Onko aineistossa joitain hyvin poikkeuksellisia mittaustuloksia?

(c) Laske piste-estimaatti loppusykkeen odotusarvojen erotukselle käsittelyjen välillä. Vastaako havaittu erotus suunnaltaan ja suuruudeltaan ennakko-odotuksia?

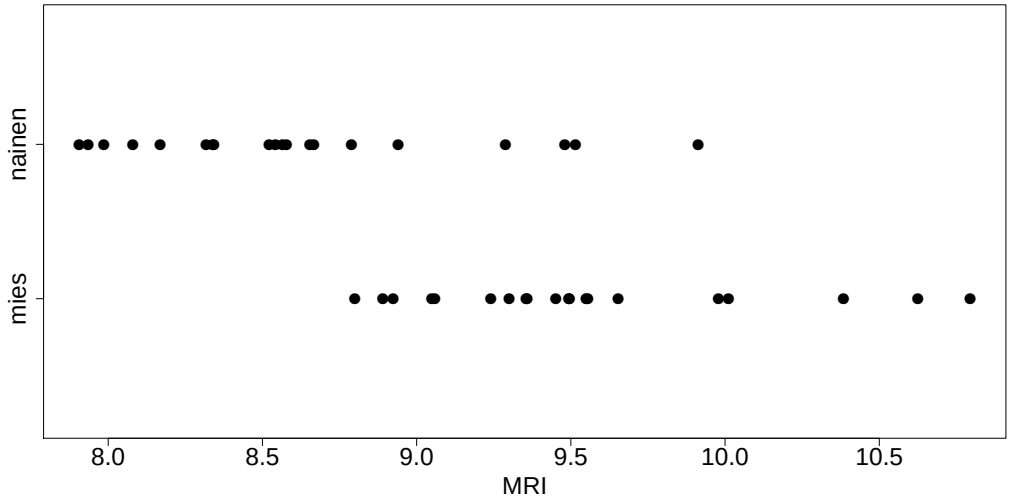
(d) Testaa nollahypoteesia, jonka mukaan loppusykkeen odotusarvoissa ei ole eroa käsittelyjen välillä: laske testisuureen arvo ja arvioi sitä vastaavaa 2-tahoista P-arvoa. Mitä informaatiota testitulos antaa? Onko havaintoaineisto sopusoinnussa nollahypoteesin kanssa? Entä ennakko-odotusten kanssa?

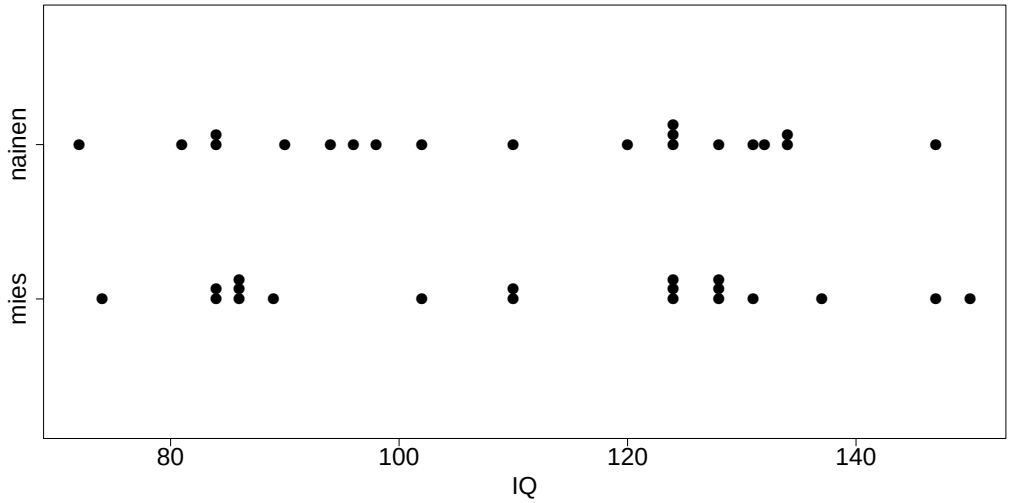
(e) Laske 95% luottamusväli loppusykkeen odotusarvojen erotukselle käsittelyjen välillä. Kuinka hyvin tulos on sopusoinnussa ennakko-oletusten kanssa?

2. Joukolta vapaaehtoisia miehiä ($n = 20$) ja naisia ($n = 20$) mitattiin heidän älykkyysosamääränsä (IQ) sekä aivojen koko (MRI) magneettikuvauslaitteen avulla (pikseleinä 18 MRI-kuvasta). Havainnot ovat datakehikossa `dats`.

Seuraavalla sivulla vertaillaan sekä graafisesti että perustunnuslukujen avulla näiden muuttujien jakaumia miesten ja naisten välillä.

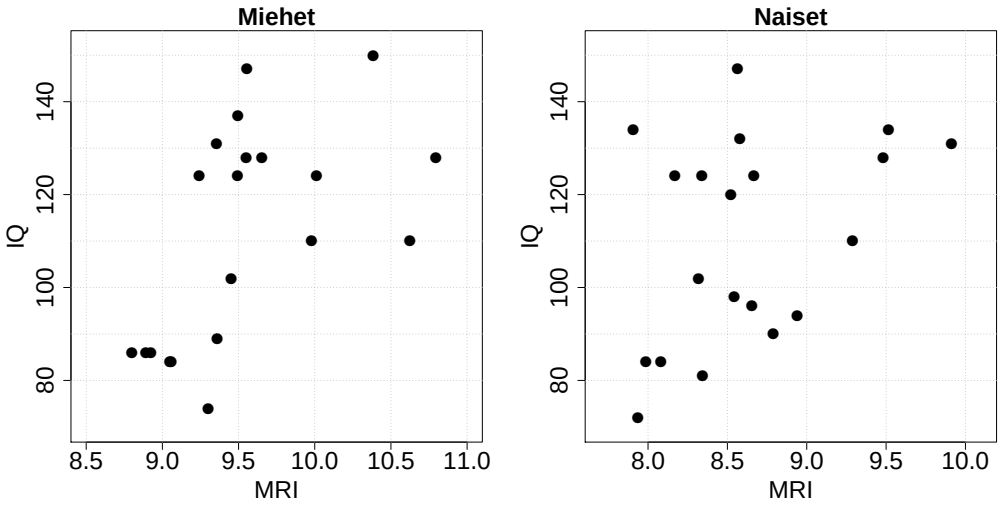
Analysoi aivojen koon jakaumien odotusarvojen erotusta miesten ja naisten populaatioiden välillä asetelmaan sopivin menetelmin ja tulkitse tulokset.






```
> with(dats, round(tapply(MRI, sukup, mean), 2))

  mies nainen
  9.55   8.63

> with(dats, round(tapply(MRI, sukup, sd), 2))

  mies nainen
  0.56   0.56

> with(dats, round(tapply(IQ, sukup, mean), 2))

  mies nainen
111.60 110.45

> with(dats, round(tapply(IQ, sukup, sd), 2))

  mies nainen
 23.54  21.95
```

3. Jatkoa edelliseen tehtävään. Älykkyysosamäärän riippuvuutta aivojen koosta sekä miehillä että naisilla havainnollistetaan seuraavissa sirontakuvioissa.



(a) Näyttääkö älykkyysosamäärä olevan yhteydessä aivojen kokoon miehillä ja/tai naisilla?

(b) Millaisen regressiomallin muodostaisit älykkyysosamäärän ja aivojen koon välille kummallakin sukupuolella erikseen? Muotoile malli ja kirjaa sen oletukset.

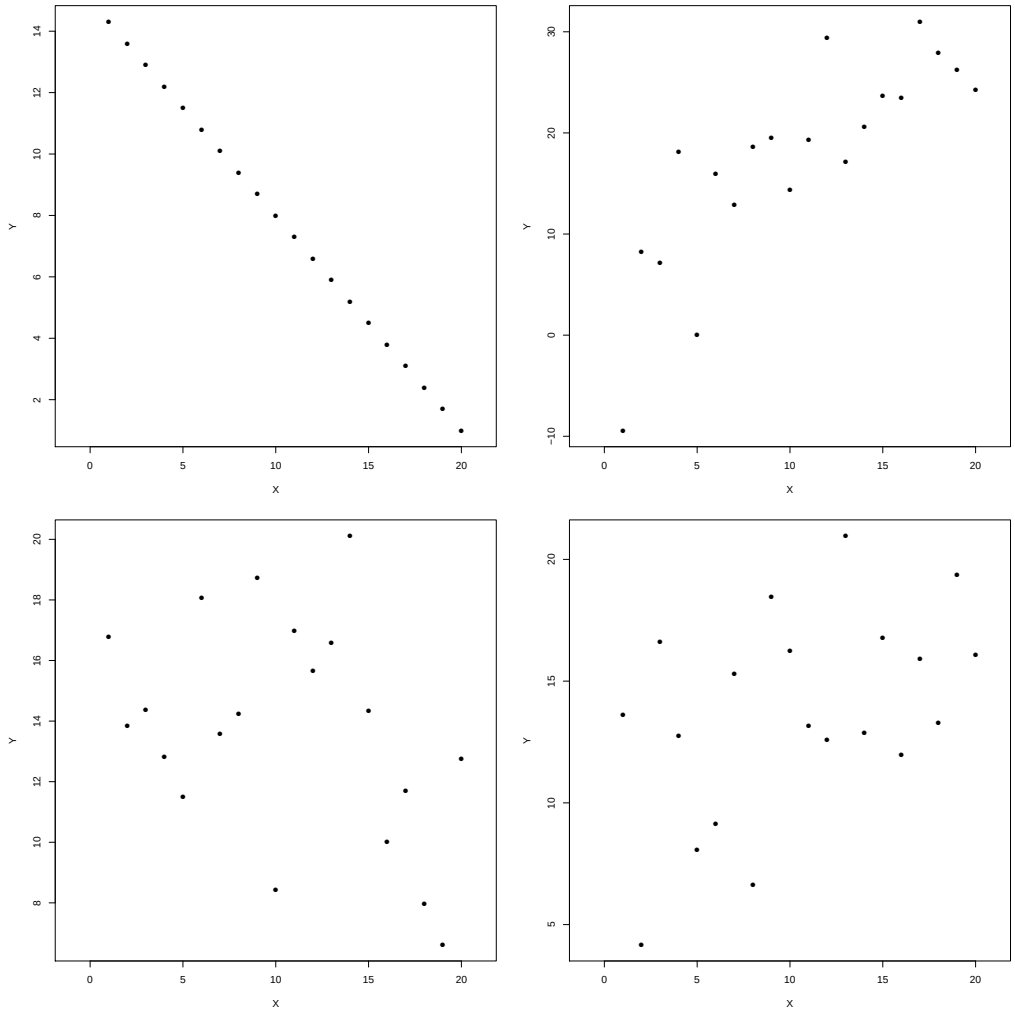
(c) Tehtävässä 2. on annettu IQ:n ja MRI:n keskiarvot ja keskihajonnat sekä miehillä että naisilla. Näiden muuttujien välinen korrelaatiokerroin oli miehillä 0.568 ja naisilla 0.396. Laske regressiokertoimien piste-estimaatit erikseen miehille ja naisille.

(d) Piirrä sekä miesten että naisten sirontakuvioon edellisessä kohdassa laskemiesi regressiokertomien piste-estimaattien mukainen sovitettu regressiosuora.

(e) Regressiosuoran kulmakertoimen estimaatin keskivirhe oli miehillä 8.17 ja naisilla 8.50. Laske kulmakertoimen 95% luottamusväli sekä miehille että naisille.

(f) Miten tulkitset tuloksia? Vaikuttaako aivojen koko älykkyysosamäärään? Jos vaikuttaa, niin onko älykkyysosamäärän odotusarvo suurempi miehillä kuin naisilla, koska miesten aivot ovat naisten aivoja keskimäärin suuremmat?

4. Seuraavassa on neljä erilaista sirontakuviota muuttujien X ja Y välillä, ja kunkin pisteparven keskelle on piirretty siihen parhaiten sopiva regressiosuora.



(a) Regressiosuorien kulmakertoimien β arvot näissä kuvioissa olivat $-0.70, 0.25, -0.37, 1.48$. Mihin kuvioon kukin näistä luvuista kuuluu?

(b) Regressiosuorien vakiokertoimien α arvot näissä kuvioissa olivat 3.5, 9.9, 15.0, 18.6. Mihin kuvioon kukin näistä luvuista kuuluu?

(c) Korrelaatiokertoimen R arvot näissä kuvioissa olivat $0.34, 0.90, -1.0, -0.63$. Mihin kuvioon kukin näistä luvuista kuuluu?