Oulun yliopiston matemaattisten tieteiden laitos/tilastotiede

JOHDATUS OTANTAMENETELMIIN (4 op), sl 2011 (Esa Läärä)

LOPPUKUULUSTELU – *aineopinnot*, ma 9.1.2012 klo 14-18 L1

Nämä kysymykset on tarkoitettu vain kurssin aineopintoina suorittaville.
Mukana saa olla tavanomainen laskin mutta ei kirjallista materiaalia. Tehtäväpaperin lopussa olevassa liitteessä on eräitä kaavoja.

1. Tampereella ilmestyvä *Aamulehti* julkaisi joulukuun alussa 2009 juuri itsenäisyyspäivän edellä Taloustutkimuksella teettämänsä kyselytutkimuksen tulokset. Kaikkiaan 1127 täysi-ikäiseltä kansalaiselta kysyttiin: "*Kuka on mielestänne oikea henkilö edustamaan Suomea EU:n huippukokouksissa?*" Vastausvaihtoehdot $d = 1, \ldots, 4$ ja niiden kannatusosuudet kyselyyn vastanneiden joukossa olivat:

1: pääministeri – 46%,

2: presidentti – 15%,

3: pääosin pääministeri mutta ulko- ja turvallisuuspoliittisissa asioissa presidentti – 34%,

4: ei osaa sanoa – 5%.

Kyselyn kohdeperusjoukon voidaan ajatella muodostuvat kaikista täysi-ikäisistä Suomen kansalaisista, ja sen koko on pyöreästi $N = 4\,000\,000$. Edelleen oletamme, että kyselyyn vastanneet ovat yksinkertainen satunnaisotos perusjoukosta. Merkitään N'_d = vaihtoehtoa d kannattavien määrä koko perusjoukossa, jossa $d = 1, \ldots, 4$.

(a) Mikä on otantasuhteen arvo, kuinka suuria ovat yksittäisten alkioiden otospainot ja miten otospaino tulkitaan?

(b) Laske ja raportoi piste-estimaatti, keskivirhe ja likimääräinen 95% luottamusväli vaihtoehtoa 1 kannattavien lukumäärälle N'_1 perusjoukossa.

2. Tee tiiviisti selkoa ositetusta otannasta: pääperiaate, käyttöaiheet, edut osittamattomaan otantaan verrattuna sekä erilaiset kiintiöintitavat.

3. Keski-Suomen läänin 32 kuntaa käsittävä perusjoukko v 1991 (ks. liite 1) on jaettu 8 eri rypääseen (muuttuja `CLU`) siten, että kussakin rypäässä on 4 maantieteellisesti toisilleen läheistä kuntaa. Näiden 8 rypään joukosta poimittiin 2 rypään otos noudattaen yksinkertaista satunnaisotantaa. Liitteessä 2 on tulostettu otokseen valittujen rypäiden kuntakohtaiset tulokset erikseen työttömien määrän (muuttuja `UE91`) ja työvoimaan kuuluvien määrän (muuttujan `LAB91`) osalta. Lisäksi näistä muttujista esitetään rypään sisäisten otosjakaumien eräitä tunnuslukuja ml. muuttujien välinen korrelaatiokerroin.

(a) Missä olosuhteissa ryväsotanta on erityisen hyödyllinen otoksen poimintatapa?

(b) Laske tämän otoksen pohjalta piste-estimaatti ja keskivirhe työttömien lukumäärälle koko läänissä.

(c) Laske otoshavainnoista piste-estimaatti työttömyysasteelle (työttömien määrän osuus työvoimaan kuuluvien määrästä) koko läänissä.

4. Kerro systemaattisesta otannasta ja sen ominaisuuksista:

(a) millä tavalla otos tätä otantatapaa noudattaen poimitaan perusjoukosta,

(b) millaiset ovat muuttujan y kokonaismäärän T_y ja keskiarvon $\bar{Y}$ sekä kahden muuttujan y ja x kokonaismäärien suhteen $R = T_y/T_x$ estimaattorit näin hankitun otoksen pohjalta,

(c) kuinka estimoinnin tarkkuuteen vaikuttavaa se, millä tavoin alkiot on järjestetty otantakehikossa?

LIITE 1: Havaintoaineisto Keski-Suomen läänin kunnista v 1991. (Lähde: Lehtonen, R., Pahkinen, E. *Practical Methods for Design and Analysis of Complex Surveys, 2nd ed.*)

Table 2.1 The *Province'91* population. Percentage unemployment (%UE) and totals of unemployed persons (UE91), labour force (LAB91), population in 1991 (POP91) and number of households (HOU85) by municipality in the province of Central Finland in 1985.

ID	LABEL	STR	CLU	%UE	UE91	LAB91	POP91	HOU85
	Urban			**12.67**	**8022**	**63 314**	**129 460**	**49 842**
1	Jyväskylä	1	1	12.20	4123	33 786	67 200	26 881
2	Jämsä	1	2	11.07	666	6016	12 907	4663
3	Jämsänkoski	1	2	13.83	528	3818	8118	3019
4	Keuruu	1	2	12.84	760	5919	12 707	4896
5	Saarijärvi	1	3	14.62	721	4930	10 774	3730
6	Suolahti	1	5	15.12	457	3022	6159	2389
7	Äänekoski	1	3	13.17	767	5823	11 595	4264
	Rural			**12.63**	**7076**	**56 011**	**125 124**	**41 911**
8	Hankasalmi	2	5	15.07	391	2594	6080	2179
9	Joutsa	2	6	9.38	194	2069	4594	1823
10	Jyväskylän mlk.	2	7	11.82	1623	13 727	29 349	9230
11	Kannonkoski	2	4	18.64	153	821	1919	726
12	Karstula	2	4	13.53	341	2521	5594	1868
13	Kinnula	2	8	13.92	129	927	2324	675
14	Kivijärvi	2	8	15.63	128	819	1972	634
15	Konginkangas	2	3	21.04	142	675	1636	556
16	Konnevesi	2	5	12.91	201	1557	3453	1215
17	Korpilahti	2	1	11.15	239	2144	5181	1793
18	Kuhmoinen	2	2	12.91	187	1448	3357	1463
19	Kyyjärvi	2	4	11.31	94	831	1977	672
20	Laukaa	2	5	12.11	874	7218	16 042	4952
21	Leivonmäki	2	6	10.65	61	573	1370	545
22	Luhanka	2	6	10.34	54	522	1153	435
23	Multia	2	7	11.24	119	1059	2375	925
24	Muurame	2	1	9.79	296	3024	6830	1853
25	Petäjävesi	2	7	15.08	262	1737	3800	1352
26	Pihtipudas	2	8	13.02	331	2543	5654	1946
27	Pylkönmäki	2	4	17.98	98	545	1266	473
28	Sumiainen	2	3	12.80	79	617	1426	485
29	Säynätsalo	2	1	10.28	166	1615	3628	1226
30	Toivakka	2	6	11.72	127	1084	2499	834
31	Uurainen	2	7	16.47	219	1330	3004	932
32	Viitasaari	2	8	14.16	568	4011	8641	3119
	Whole province			**12.65**	**15 098**	**119 325**	**254 584**	**91 753**

Sources: Statistics Finland: Population Census 1985. Statistics Finland (1992): Statistical Yearbook of Finland. Volume 87. Ministry of Labour of Finland (1991): Employment Service Statistics. November 30, 1991.

LIITE 2: Keski-Suomen läänin kunnista poimittu kahden rypään otos: kunnittaiset havainnot kummassakin rypäässä sekä rypäiden sisäisiä tunnuslukuja työttömien määrän (UE91) ja työvoimaan kuuluvien lukumäärän (LAB91) otosjakaumista ml. näiden muuttujien välinen korrelaatiokerroin (Pearson Correlation Coefficient).

Obs	CLU	ID	Municipality	UE91	LAB91
1	2	2	Jämsä	666	6016
2	2	3	Jämsänkoski	528	3818
3	2	4	Keuruu	760	5919
4	2	18	Kuhmoinen	187	1448
5	7	10	Jyväskmlk	1623	13727
6	7	23	Multia	119	1059
7	7	25	Petäjävesi	262	1737
8	7	31	Uurainen	219	1330

---------- CLU=2 ----------

Simple Statistics

Variable	N	Mean	Std Dev	Minimum	Maximum
UE91	4	535	251	187	760
LAB91	4	4300	2155	1448	6016

Pearson Correlation Coefficients, N = 4

	UE91	LAB91
UE91	1.00000	0.97689
LAB91	0.97689	1.00000

---------- CLU=7 ----------

Simple Statistics

Variable	N	Mean	Std Dev	Minimum	Maximum
UE91	4	556	714	119	1623
LAB91	4	4463	6182	1059	13727

Pearson Correlation Coefficients, N = 4

	UE91	LAB91
UE91	1.00000	0.99902
LAB91	0.99902	1.00000

LIITE 3: Eräitä merkintöjä ja kaavoja

Poimintavektorin $I = (I_1, \ldots, I_N)$ momentit $(i, j = 1, \ldots, N)$:

$$\tau_i = \mathbb{E}(I_i), \qquad \Delta_{ii} = \mathrm{var}(I_i) = \mathbb{E}(I_i^2) - [\mathbb{E}(I_i)]^2,$$
$$\tau_{ij} = \mathbb{E}(I_i I_j), \quad \Delta_{ij} = \mathrm{cov}(I_i, I_j) = \mathbb{E}(I_i I_j) - \mathbb{E}(I_i)\mathbb{E}(I_j).$$

Lavennetut y:n arvot: $\check{y}_i = y_i/\tau_i$ ja kovarianssit: $\check{\Delta}_{ij} = \Delta_{ij}/\tau_{ij}$.

Satunnaisvektorin $X = (X_1, \ldots, X_m)$ koordinaattimuuttujien lineaarikombinaation odotusarvo ja varianssi, kun $c_1, \ldots, c_m$ ovat reaalivakioita:

$$\mathbb{E}\left(\sum_{i=1}^{m} c_i X_i\right) = \sum_{i=1}^{m} c_i \mathbb{E}(X_i),$$
$$\mathrm{var}\left(\sum_{i=1}^{m} c_i X_i\right) = \sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j \mathrm{cov}(X_i, X_j) = \sum_{i=1}^{m} c_i^2 \mathrm{var}(X_i) + 2\sum_{i=1}^{m}\sum_{j>i} c_i c_j \mathrm{cov}(X_i, X_j)$$

Olkoot X ja Y kaksi satunnaismuuttujaa, joiden odotusarvot ovat μ_X ja μ_Y, varianssit σ_X^2 ja σ_Y^2, ja keskinäinen kovarianssi σ_{XY}. Olkoon edelleen $Z = g(X, Y)$ pisteessä (μ_X, μ_Y) vähintään kahdesti derivoituva muunnos. Muunnoksen odotusarvoa ja varianssia voidaan approksimoida

$$\mathbb{E}[g(X,Y)] \approx g(\mu_X, \mu_Y),$$
$$\mathrm{var}[g(X,Y)] \approx \left(\frac{\partial g}{\partial X}\right)^2 \sigma_X^2 + \left(\frac{\partial g}{\partial Y}\right)^2 \sigma_Y^2 + 2\left(\frac{\partial g}{\partial X}\right)\left(\frac{\partial g}{\partial Y}\right)\sigma_{XY},$$

jossa osittaisderivaatat evaluoidaan pisteessä (μ_X, μ_Y).

Suhdeluvun $R = T_y/T_x$ estimaattorin $\widehat{R} = \widehat{T}_y/\widehat{T}_x$ 1. kertaluvun Taylorin kehitelmä:

$$\widehat{R} \approx R + \frac{1}{T_x}(\widehat{T}_y - R\widehat{T}_x)$$

Lause 5.1. Muuttujan y kokonaismäärän $T = \sum_U y_i$ harhaton estimaattori on $\widehat{T} = \sum_U \check{y}_i I_i$, ja tämän varianssi sekä varianssin harhaton estimaattori ovat

$$\mathrm{var}(\widehat{T}) = \sum_{i=1}^{N}\sum_{j=1}^{N} \Delta_{ij}\check{y}_i\check{y}_j, \qquad \widehat{\mathrm{var}}(\widehat{T}) = \sum_{i=1}^{N}\sum_{j=1}^{N} \check{\Delta}_{ij}\check{y}_i\check{y}_j I_i I_j.$$

Lause 5.2. Kun otoskoko n on kiinteä, niin $\widehat{T}$:n varianssi $\mathrm{var}(\widehat{T})$ voidaan kirjoittaa allaolevassa muodossa, ja $\mathrm{var}(\widehat{T})$:n harhaton SYG-estimaattori on seuraavan kaavan mukainen

$$\mathrm{var}(\widehat{T}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \Delta_{ij}(\check{y}_i - \check{y}_j)^2, \qquad \widehat{\mathrm{var}}_{\mathrm{SYG}}(\widehat{T}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \check{\Delta}_{ij}(\check{y}_i - \check{y}_j)^2 I_i I_j$$

Multinomiasetelma: $I = (I_1, \ldots, I_N) \sim \mathrm{Multin}(n, p_1, p_2, \ldots, p_N)$, jossa $p_1 + \cdots + p_N = 1$, pistetodennäköisyydet (kun $k_i \in \mathbb{N}$ ja $k_1 + \cdots + k_N = n$) ja momentit:

$$\mathbb{P}\{I_1 = k_1, \ldots, I_N = k_N\} = \frac{n!}{k_1! \cdots k_N!} p_1^{k_1} \cdots p_N^{k_N},$$
$$\tau_i = np_i,$$
$$\Delta_{ij} = -np_i p_j, \quad \tau_{ij} = n(n-1)p_i p_j, \; i \neq j,$$
$$\Delta_{ii} = np_i(1 - p_i), \quad \tau_{ii} = np_i(1 - p_i + np_i).$$