VAASAN KORKEAKOULUN JULKAISUJA

TUTKIMUKSIA No 96

Tilastotiede 13

Erkki Latosaari - Ilkka Virtanen

ENTROPIA HOMOGEENISUUSINDEKSINÄ RYHMITTELYANALYYSISSÄ

SUMMARY

Entropy as a measure of homogeneity
in categorical grouping analysis

Vaasa 1983

SISÄLLYSLUETTELO

ABSTRACT

Latosaari, Erkki & Virtanen, Ilkka (1983). Entropia homogeenisyysindeksinä ryhmittelyanalyysissä (Entropy as a measure of homogeneity in categorical grouping analysis). Proceedings of the University of Vaasa. Research Papers No 96, 44 p.

The paper deals with the concept of Shannon's entropy from the point of view of statistics. Entropy is considered as a measure of dispersion for a categorical variable. Of special interest in the paper is the case where the classes or categories of the variable have been aggregated to form homogeneous (with respect to the class frequencies) groups. The total entropy of the variable is divided into two components, the entropy between the groups and the entropy within the groups. This division forms the basis for analyzing the homogeneity of the aggregated groups. Further, an entropy-based test statistic, viz. Kullback's information statistic, is introduced to carry out homogeneity tests for the groups in the case of sample data. The grouping procedure is illustrated with an application to the finnish representative elections.

Correspondence to Professor Ilkka Virtanen, University of Vaasa, School of Business Studies, Raastuvankatu 31, SF-65100 Vaasa 10, Finland.

Key words: categorical variables, entropy decomposition, grouping analysis, information statistic, measure of homogeneity.

1. JOHDANTO

Entropian käsite on alkuaan peräisin fysiikasta, jossa entropia liittyy termodynaamisten systeemien tilojen muutoksiin. Termodynamiikan II pääsäännön eräs esitysmuoto on lausuttu entropiamuodossa: fysikaalisessa systeemissä vain sellaiset tilamuutokset ovat mahdollisia, joissa kokonaisentropia kasvaa. Siirtymällä systeemin makrotason tiloista hiukkastasolle saadaan entropiasta mittari mikrotason epäjärjestykselle.

Entropia on keskeinen käsite myös informaatioteoriassa, erityisesti erilaisten tiedonsiirtojärjestelmien yhteydessä. Informaatioteoriasta on peräisin myös se entropian matemaattinen esitysmuoto, jota sittemmin on laajamittaisesti sovellettu eri tieteenaloilla. Shannonin mukaan (Shannon 1948) diskreettiin todennäköisyysjakaumaan ($p_1, p_2, \ldots, p_n$) liittyy sitä suurempi epävarmuus, mitä laakeampi ja tasaisempi jakauma on. Jakauman Shannonilla muodostivat sanoman lähetyksessä käytetyt merkit ja epävarmuudella tarkoitettiin epävarmuutta vastaanotetusta sanomasta suhteessa lähetettäväksi mahdollisten sanomien määrään. Mitta tälle epävarmuudelle, nimeltään jakauman entropia, määriteltiin lausekkeena

$$(1.1) \qquad S(p_1, p_2, \ldots, p_n) = - k \sum_{i=1}^{n} p_i \log p_i \ .$$

Entropia-käsitteen määrittelyä, tulkintoja ja käyttöä edellä mainituilla pioneerialoilla on lähemmin esitellyt mm. Latosaari (1983).

Runsaan vuosikymmenen ajan on entropian sekä termodynaamisia että informaatioteoreettisia analogioita ja tulkintoja sovellettu myös yhteiskunta- ja taloustieteisiin. Georgescu-Roegen (1967: 92-129) on soveltanut entropian termodynaamista lähestymistapaa taloudellisten prosessien yleiseen kuvailuun. Theil (1969) voidaan mainita pioneerina liiketaloustieteisiin suuntautuneiden sovellutusten alueella. Wilson toi entropiakäsitteen varsin varhain talousmaatieteeseen alue- ja kuljetussysteemien tarkastelun yhteydessä (Wilson 1970a). Hän on myöskin tarkastellut entropiaa mallityöskentelyn ja operaatiotutkimuksen yleisenä

työvälineenä (Wilson 1970b). Latosaari on tutkielmassaan käsitellyt laajahkosti näitä entropia-käsitteen taloustieteellisiä sovellutuksia (Latosaari 1983).

Informaatioteoriassa käytetty entropian käsite on ollut sysäyksenä suureen käytölle myös tilastotieteessä. Kullback on kirjoittanut hyvän oppikirjan tilastotieteen ja informaatioteorian yhteyksistä yleisemminkin tarkasteltuna (Kullback 1959).

Perinteisesti entropiaa on tilastotieteessä käytetty nominaaliasteikon hajontalukuna. Nominaaliasteikolla esitetyn jakauman hajonnalla on varsin läheinen yhteys Shannonin epävarmuus-käsitteeseen. Jakauman hajontaa kuvaavan entropian tulkitseminen matemaattisesti muuttujan ekvivalenssiluokkiin liittyvän epävarmuuden odotusarvoksi tuo mielenkiintoisella tavalla esiin entropian määritelmällisen analogian välimatka- tai suhdeasteikon tasoisen muuttujan varianssin (tai keskihajonnan) kanssa (Astola ja Virtanen 1981: 12-13).

Sekä Kullback (1959) että Theil (1969) ovat osoittaneet entropiaan perustuvilla suureilla olevan käyttöä myös kaksi- ja useampiulotteisten frekvenssiaineistojen (kontingenssitaulujen) yhteydessä. Entropia toimii tällöin taulukon muuttujien välillä esiintyvän tilastollisen riippuvuuden ilmaisijana ja mittarina. Preuss (1980) on vienyt tämän entropia-käsitteeseen perustuvan riippuvuusanalyysin tunnuslukuasteelle saakka. Astola ja Virtanen (1981, 1982, 1983) ovat kehittäneet analyysia niin, että käyttöön on saatu korrelaatiokertoimen hyvyysominaisuudet toteuttava riippuvuusluku, entropiakorrelaatiokerroin.

Tässä tutkimuksessa entropiaa tarkastellaan nimenomaan tilastotieteellisenä suureena, yksiulotteisen luokitusasteikon tasoisen jakauman hajonnan mittana. Kohteena on erityisesti ryhmittelyanalyysi, jossa pyritään selvittämään luokista luonnollisella tavalla muodostuneiden tai tietyillä periaatteilla muodostettujen aggregoitujen ryhmien sisäistä homogeenisuutta ja ryhmien välistä heterogeenisuutta. Tutkimuksessa Astola ja Virtanen (1981) esitetty määritelmällinen ja tulkinnallinen analogia käsiteparin entropia - varianssi välillä saa tässä luontevan jatkon: jakauman kokonaisentropialla ja sen ryhmäjaon synnyttämillä komponenteilla, ryhmien sisäisellä entropialla ja ryhmien välisellä entropialla on selvät vastineet varianssianalyysin vastaavissa varianssisuureissa. Informaatiokertoimien (entropian johdannaisia) käyttöä ryhmittelyanalyysissä on aiemmin tutkinut mm. Pihlajarinne (1979). Tässä käytettävä lähestymistapa ja menetelmät ovat kuitenkin täysin poikkeavat Pihlajarinteen vastaavista.

Sovellutuksena tutkimuksessa tarkastellaan Suomen eduskuntapuolueiden kansanedustajain vaaleissa saavuttaman kannatusosuuden ja saatujen kansanedustajien lukumäärän jakaumia vv. 1970-1983. Suoritettava ryhmittelyanalyysi vahvistaa yleisesti omaksutun käsityksen, jonka mukaan eduskuntaan on muodostunut sekä suhteellisen vakiintunut suurten puolueitten ryhmä että hieman vähemmän vakiintunut pienten puolueitten ryhmä. Tarkasteltavaan ajanjakson alkupuolella tämä ryhmäjako edelleen selkiintyi ja vahvistui kunnes viimeisissä vaaleissa kehitys pysähtyi ja muutti suuntaansakin. Edelleen analyysi tuottaa numeerisen indeksin sekä ryhmien sisäiselle homogeenisuudelle että niiden väliselle heterogeenisuudelle.

2. ENTROPIA RYHMITTELYANALYYSISSÄ

2.1. Entropian perusominaisuudet

Tarkastellaan kvalitatiivista, ts. luokitus- tai järjestysasteikon tasoista (satunnais-)muuttujaa X, jonka jakauma on seuraava:

Ekvivalenssiluokka	E_1	E_2	$\cdots$	E_i	$\cdots$	E_n
Todennäköisyys (suhteellinen frekvenssi)	p_1	p_2	$\cdots$	p_i	$\cdots$	p_n

Aidolle kvalitatiiviselle muuttujalle, jolle luokkien symbolit ("muuttujan arvot") voivat olla täysin mielivaltaisia, on kaikki tunnusluvut määriteltävä pelkästään luokkafrekvenssien tai -todennäköisyyksien mukaan. Hajontalukuna käytetty entropia

$$(2.1) \qquad H = -\sum_{i=1}^{n} p_i \log_2 p_i$$

täyttää nyt selvästikin tämän vaatimuksen. Kun entropian saamat arvot lisäksi hyvin vastaavat intuitiivista käsitystä jakauman keskittyneisyydestä tai hajaantuneisuudesta, on sen käyttö kvalitatiivisen muuttujan hajontalukuna kaikin puolin perusteltu. Tärkeimmät äsken mainituista ominaisuuksista ovat: H on ei-negatiivinen, $H = 0$ tarkalleen silloin, kun jakauma on keskittynyt yhteen luokkaan (tarkalleen yksi $p_i = 1$), H saavuttaa suurimman arvonsa ($\log_2 n$) tasajakauman tapauksessa, ts. kun $p_1 = p_2 = \cdots = p_n = 1/n$ (esim. Astola ja Virtanen 1981: 7-10).

Määritelmistä (1.1) ja (2.1) nähdään, että (1.1):ssä esiintyvä vakio k kiinnitetään tavallisesti käytetyn logaritmijärjestelmän kantaluvun valinnalla. Valitsemalla samanaikaisesti $k = 1$ ja kantaluvuksi 2 päästään määritelmään (2.1), jolloin entropian yksikköä kutsutaan yleisesti bitiksi. Kantaluvun valinta ei sinänsä ole tärkeä, koska entropian absoluuttisella arvolla ei yksinään ole suurtakaan merkitystä. Tämä puolestaan johtuu entropian saamien arvojen skaalaamattomuudesta, arvojen yläraja riippuu käytetystä luokkajaosta ($=\log_2 n$). Kiinteällä luokkien lukumäärällä voidaankin siirtyä ns. suhteelliseen entropiaan

$$(2.2) \qquad \tilde{H} = \frac{H}{H_{max}} = -\frac{\sum_{i=1}^{n} p_i \log_2 p_i}{\log_2 n},$$

joka saa arvoja suljetulta väliltä $[0,1]$. Määritelmästä (2.2) nähdään, että suhteellinen entropia $\tilde{H}$ on käytetystä logaritmijärjestelmästä ja luokkien lukumäärästä riippumaton.

Shannonin esittämä tulkinta entropialle epävarmuuden mittana on myös todennäköisyysteoreettisesti perusteltavissa. Tarkastellaan satunnaismuuttujaa $H = H(X)$, joka saa arvon $\eta_i = \log (1/p_i)$ kun muuttujan X arvo x_i kuuluu ekvivalenssiluokkaan E_i. Arvo η_i kuvaa selvästi luokan E_i esiintymiseen liittyvää epävarmuutta: η_i on sitä suurempi (luokka E_i epävarmempi) mitä pienempi luokan esiintymistodennäköisyys p_i on. Lauseke (2.1) voidaan nyt saattaa muotoon (tästä eteenpäin jätetään logaritmijärjestelmän kantaluku 2 merkitsemättä)

(2.3) $$H = -\sum_{i=1}^{n} p_i \log p_i$$

$$= \sum_{i=1}^{n} p_i \log (1/p_i)$$

$$= \sum_{i=1}^{n} p_i \eta_i$$

$$= E \{ H \},$$

ts. entropia H on lausuttu satunnaismuuttujan H eli muuttujaan X liittyvän epävarmuuden odotusarvona. Vastaavanlainen odotusarvotulkinta liittyy moniulotteisenkin jakauman yhteisentropiaan (Astola ja Virtanen 1981: 13-14).

Entistä mielenkiintoisemmaksi esitysmuodon (2.3) tekee siitä johdettavissa oleva analogia kvalitatiivisen muuttujan entropian ja kvantitattivisen muuttujan varianssin välille. Sillä onhan varianssi

(2.4) $$V = E \{(X - E\{X\})^2\} = E \{ D \}$$

niinikään määritelty odotusarvona, perustana oleva satunnaismuuttuja on nyt vain muuttujan X saamiin arvoihin liittyvä epätarkkuus $D = D(X) = (X - E\{X\})^2$, jolloin epätarkkuudella ymmärretään odotusarvon suhteen laskettua neliöpoikkeamaa. Analogia (2.3):n ja (2.4):n välillä on ilmeinen, luonteidensa mukaisesti kvantitatiivisen muuttujan epätarkkuutta vain mitataan muuttujan saamien arvojen avulla ja kvalitatiiviisen muuttujan epävarmuutta arvoihin liittyvien todennäköisyyksien avulla.

2.2. Kokonaisentropia ja sen komponentit

Edellä kuvattu analogia entropian ja varianssin välillä johtaa luontevasti kysymykseen siitä, onko mahdollista suorittaa varianssianalyysin tapaista tarkastelua myös entropia-suureen avulla. Osoittautuukin, että mikäli muuttujan luokista muodostetaan aggregoituja ryhmiä, niin kokonaisentropia on esitettävissä varianssin tapaan summana ryhmien sisäisestä ja ryhmien välisestä entropiasta. Tätä entropian ominaisuutta onkin hyödynnetty eräissä taloudellisissa sovellutuksissa (Theil 1969: 459-469, Kettunen 1973: 52-70, Horowitz ja Horowitz 1976: 121-122, Walker, Stowe ja Moriarty 1979: 173-186).

Oletetaan nyt seuraavassa, että jakson 2.1 jakaumassa luokat $E_1,\ldots,E_{g_1}$ muodostavat ryhmän G_1, luokat $E_{g_1+1},\ldots,E_{g_1+g_2}$ ryhmän G_2 ,..., ja luokat $E_{g_1+\ldots+g_{k-1}+1}$,..., $E_{g_1+\ldots+g_k}$ ryhmän G_k , ts. näitä aggregoituja ryhmiä on k kpl ja ryhmässä G_j on g_j alkuperäistä luokkaa yhdistettynä ($j=1,2,\ldots,k$). Merkitään edelleen ryhmän G_j todennäköisyyttä q_j:llä, jolloin (määrittelemällä $q_0 = 0$)

(2.5) $$q_j = \sum_{r=1}^{g_j} p_{g_0+\ldots+g_{j-1}+r} , \quad j=1,2,\ldots,k.$$

Ryhmien avulla lausuttuna jakauma on nyt edellä kuvatuin symbolein seuraava:

Ryhmä	G_1	G_2	...	G_j	...	G_k
Todennäköisyys	q_1	q_2	...	q_j	...	q_k.

Tämän aggregoidun jakauman entropia, eli alkuperäisen jakauman kannalta tarkasteltuna ryhmien välinen entropia (ryhmäjaon "selittämä" entropia), on

$$(2.6) \qquad H_B = -\sum_{j=1}^{k} q_j \log q_j$$

eli luokkiin liittyvin symbolein lausuttuna

$$(2.7) \qquad H_B = -\sum_{j=1}^{k} \sum_{r=1}^{g_j} p_{g_0+\ldots+g_{j-1}+r} \log \Big(\sum_{r=1}^{g_j} p_{g_0+\ldots+g_{j-1}+r} \Big).$$

Tarkastellaan seuraavaksi ryhmän G_j $(j=1,2,\ldots,k)$ sisäistä entropiaa. Jakauma ryhmän sisällä on nyt

Luokka $\quad E_{g_0+\ldots+g_{j-1}+1} \cdots E_{g_0+\ldots+g_{j-1}+r} \cdots E_{g_0+\ldots+g_j}$

Todennäköisyys $\quad p'_{g_0+\ldots+g_{j-1}+1} \cdots p'_{g_0+\ldots+g_{j-1}+r} \cdots p'_{g_0+\ldots+g_j}$

missä

$$(2.8) \qquad p'_{g_0+\ldots+g_{j-1}+r} = \frac{1}{q_j} p_{g_0+\ldots+g_{j-1}+r}$$

$$= \frac{p_{g_0+\ldots+g_{j-1}+r}}{\sum_{r=1}^{g_j} p_{g_0+\ldots+g_{j-1}+r}} \quad , j=1,\ldots,k, r=1,\ldots,g_j.$$

Ryhmän G_j sisäinen entropia $H_W^{(j)}$ on näin

$$(2.9) \qquad H_W^{(j)} = -\sum_{r=1}^{g_j} p'_{g_0+\ldots+g_{j-1}+r} \log p'_{g_0+\ldots+g_{j-1}+r} .$$

Määritellään seuraavaksi ryhmien sisäinen entropia koko jakaumassa (ryhmäjaon kannalta "selittämätön" entropia). Luonnollinen määrittely tälle sisäiselle entropialle on ryhmäkohtaisten sisäisten entropioiden $H_W^{(j)}$ odotusarvo:

$$(2.10) \qquad H_W = \sum_{j=1}^{k} q_j H_W^{(j)} .$$

Kun nyt muodostetaan ryhmien välisen ja ryhmien sisäisen entropian summa

$$(2.11) \quad H_B + H_W = -\sum_{j=1}^{k} q_j \log q_j + \sum_{j=1}^{k} q_j H_W^{(j)}$$

$$= -\sum_{j=1}^{k} \left[q_j \log q_j - q_j H_W^{(j)} \right]$$

$$= -\sum_{j=1}^{k} q_j \left[\log q_j + \sum_{r=1}^{g_j} p'_{g_0+\ldots+g_{j-1}+r} \log p'_{g_0+\ldots+g_{j-1}+r} \right]$$

$$= -\sum_{j=1}^{k} \sum_{r=1}^{g_j} p_{g_0+\ldots+g_{j-1}+r} \left[\log \sum_{r=1}^{g_j} p_{g_0+\ldots+g_{j-1}+r} + \right.$$

$$\left. \log \frac{p_{g_0+\ldots+g_{j-1}+r}}{\sum_{r=1}^{g_j} p_{g_0+\ldots+g_{j-1}+r}} \right]$$

$$= -\sum_{j=1}^{k} \sum_{r=1}^{g_j} p_{g_0+\ldots+g_{j-1}+r} \log p_{g_0+\ldots+g_{j-1}+r}$$

$$= -\sum_{i=1}^{n} p_i \log p_i = H_T ,$$

huomataan, että saadaan tulokseksi kokonaisentropia ("selitettävä" entropia). On siis saatu johdetuksi tärkeä tulos kokonaisentropian jakautumisesta kahteen komponenttiin, ryhmien väliseen entropiaan ja ryhmien sisäiseen entropiaan:

$$H_T = H_B + H_W \,. \tag{2.12}$$

Kokonaisentropiasta on sitä suurempi osa ryhmien sisäistä entropiaa, mitä homogeenisempia (luokkakoon suhteen) ryhmät sisäisesti ovat. Äärітapauksen muodostavat sisäisesti tasajakautuneet ryhmät G_j. Tällöin on

$$H_{W_{max}}^{(j)} = \log g_j \,, \quad j=1,2,\ldots,k \tag{2.13}$$

ja (pitäen ryhmätason jakaumaa kiinnitettynä)

$$H_{W_{max}} = \sum_{j=1}^{k} q_j \log g_j \,. \tag{2.14}$$

Toisen äärimmäisyyden muodostaa jakauma, jossa kunkin ryhmän sisällä todennäköisyys on keskittynyt tarkalleen yhteen luokkaan. Tällöin taas on

$$H_W^{(j)} = 0, \quad j=1,2,\ldots,k, \tag{2.15}$$

josta edelleen

$$H_W = 0, \tag{2.16}$$

ts. kokonaisentropia on kokonaisuudessaan ryhmien välistä entropiaa. Kyseessä on ryhmittelyn kannalta tavallaan surkastunut tapaus, kukin ryhmä samaistuu yhteen sen luokkaan.

Ryhmien välinen entropia H_B puolestaan on maksimissaan, kun ryhmätason jakauma $(q_1, q_2, \ldots, q_k)$ on tasainen, ts. kun $q_j = 1/k$, $j=1,2,\ldots,k$. Tällöin on

$$H_{B_{max}} = \log k \,. \tag{2.17}$$

Ryhmien välinen entropia on minimissään, kun jollekin ryhmälle G_m pätee $q_m = 1$ ja muille ryhmille $q_j = 0$, $j \neq m$. Tällöin on

$$H_B = 0 \,, \tag{2.18}$$

ts. kokonaisentropia on kokonaisuudessaan ryhmien sisäistä (itseasiassa ryhmän G_m sisäistä) entropiaa. Kyseessä on jälleen ryhmittelyn kannalta surkastunut tapaus.

Edellä on tarkasteltu kokonaisentropian ja sen komponenttien absoluuttisia arvoja ja niiden vaihtelumahdollisuuksia. Jakson 2.1 tapaan voidaan laskea myös suhteelliset, välille [0,1] skaalatut, entropiat. Kiinteällä ryhmittelyllä, kuten edellä on kaiken aikaan oletettu, entropioiden H_T ja H_B maksimiarvot ovat selvät: luokkien ja ryhmien lukumäärien logaritmit. Entropian H_W kohdalla maksimiarvo on sen sijaan ongelmallisempi. Sillä vaikka kunkin ryhmän sisällä sisäisen entropian maksimiarvo riippuu vain ryhmän luokkien lukumäärästä, vrt. (2,13), niin sisäisen entropian kokonaismäärän maksimiarvoon (2,14) vaikuttaa myös ryhmätason jakauma $(q_1, q_2, \ldots, q_k)$, jolla sinänsä ei ole mitään tekemistä ryhmien sisäisen homogeenisuuden kanssa. Ongelmasta päästään, kun määritellään suhteellinen sisäinen entropia toteutuneena sisäisenä entropiana suhteessa siihen

sisäisen entropian maksimiarvoon, jonka ryhmäjako toteutuneine ryhmätodennäköisyyksineen mahdollistaa.

Näin saadaan kokonaisentropialle ja sen komponenteille suhteelliset arvot

$$(2.19) \qquad \tilde{H}_T = \frac{H_T}{H_{T_{max}}} = \frac{H_T}{\log n}$$

$$(2.20) \qquad \tilde{H}_B = \frac{H_B}{H_{B_{max}}} = \frac{H_B}{\log k}$$

$$(2.21) \qquad \tilde{H}_W = \frac{H_W}{H_{W_{max}}} = \frac{H_W}{\sum_{j=1}^{k} q_j \log q_j} .$$

Erääksi ryhmittelyä kuvaavaksi mittariksi voidaan ajatella myös ryhmittelyn homogeenisuusastetta (tai selitysastetta):

$$(2.22) \qquad \eta^2 = 1 - \frac{H_B}{H_T} = \frac{H_W}{H_T} .$$

Koska on voimassa sekä $0 \leq H_B \leq H_T$ että $0 \leq H_W \leq H_T$, on aina $0 \leq \eta^2 \leq 1$. Selitysaste $\eta^2 = 0$, kun $H_W = 0$, ts. kun ryhmittely ei synnytä tai "selitä" homogeenisuutta lainkaan (ryhmät sisäisesti mahdollisimman vähän tasajakautuneita). Tapauksessa $H_B = 0$ on $\eta^2 = 1$, tällöin ryhmäjakauma on mahdollisimman kaukana tasajakaumasta, kokonaisentropia "selittyy" kokonaisuudessaan ryhmien sisäisellä entropialla eli ryhmien sisäisellä homogeenisuudella. Suureella η^2 (tai sen komplementtisuureella) ei kuitenkaan ole, toisin kuin esim. varianssianalyysin vastaavalla suureella, erityisen suurta merkitystä. Tämä näkyy mm. päätearvojen

$\eta^2 = 0$ ja $\eta^2 = 1$ monikäsitteisyydestä: arvo $\eta^2 = 0$ ($H_W = 0$) on täysin ryhmätason jakaumasta (H_B :stä) riippumaton, vastaavasti on arvo $\eta^2 = 1$ ($H_B = 0$) täysin ryhmän G_m (vrt. merkintä edellä) sisäisestä jakaumasta ($H_W^{(m)}$:stä ja samalla H_W :stä) riippumaton.

Esimerkki 2.1. Tarkastellaan numeroesimerkkinä edellä esitetystä entropia-analyysistä puolueiden kannatuksen jakaumaa vuoden 1979 kansanedustajain vaaleissa. Puolueiden kannatusosuudet olivat tällöin (Mitä-Missä-Milloin 1980: 146-147)

Puolue	SDP	KOK	SKDL	KESK	SKL	SMP	RKP	LKP	MUUT	YHT.
Kannatusosuus	0.239	0.217	0.179	0.173	0.048	0.046	0.045	0.037	0.016	1.000

Entropiaan perustuvassa ryhmittelyanalyysissä muuttujan (tässä puolue) ekvivalenssiluokilla tai arvoilla ei ole varsinaista sisällöllistä merkitystä, analyysi voi käyttää hyödykseen ainoastaan luokkien suhteellisia frekvenssejä. Sen kaltaiset ryhmittelyt kuin "vasemmisto", "keskusta" ja "oikeisto" tai "sosialistiset puolueet" ja "ei-sosialistiset puolueet" eivät siten nyt ole mielekkäitä, koska ne perustuvat nimenomaan luokkien sisällölliseen merkitykseen. Suhteellisiin frekvensseihin perustuva on sen sijaan ryhmittely "suuret puolueet" ja "pienet puolueet". Esimerkkinä olevista v. 1979 vaalituloksistakin nämä ryhmät erottuvat varsin selvästi. Ensimmäiseen ryhmään kuuluvat SDP, KOK, SKDL ja KESK sekä jälkimmäiseen SKL, SMP, RKP, LKP ja yhdistetty luokka MUUT. Tämän puolueiden kannatuksen suuruuteen perustuvan ryhmittelyn ajallista kehitystä ja yleistä mielekkyyttä on

käsitelty tarkemmin luvussa 3.

Puoluejakauman kokonaisentropia v. 1979 vaaleissa oli

$$(2.23) \qquad H_T = -(0.239 \ln 0.239 + \cdots + 0.016 \ln 0.016) = 1.900.$$

Yhtälössä (2.23) ja numeroesimerkeissä yleensäkin on laskuteknisistä syistä käytetty 2-järjestelmän logaritmien sijasta luonnollisia logaritmeja. Tämä merkitsee vain biteissä lausuttujen arvojen kertomista vakiolla ln 2 eli siirtymistä yksiköissä nitteihin. Suhteellisiin arvoihin kantaluvulla ei ole vaikutusta. Koska muuttujan luokkia on kaikkiaan 9, oli jakauman suhteellinen entropia

$$(2.24) \qquad \tilde{H}_T = \frac{1.900}{\ln 9} = 0.865 .$$

Tulos merkitsee, että puoluekannatuksen jakauma v. 1979 vaaleissa oli " 86.5 -prosenttisesti tasajakautunut" tai vaihtoehtoisesti " 13.5 -prosenttisesti keskittynyt" yksipuoluesuuntaan.

Ryhmäjakoon "suuret puolueet" (80.8 %) - "pienet puolueet" (19.2 %) perustuvan ryhmätason jakauman entropia (= entropia ryhmien välillä) oli v. 1979 vastaavasti

$$(2.25) \qquad H_B = -(0.808 \ln 0.808 + 0.192 \ln 0.192) = 0.489 .$$

Suhteellisena arvona saadaan tästä (k = 2)

$$(2.26) \qquad \tilde{H}_B = \frac{0.489}{\ln 2} = 0.706 ,$$

ts. kannatus ryhmien kesken on vielä " 70.6 -prosenttisesti tasajakautunutta" ja vain " 29.4 -prosenttisesti keskittynyt" suurten puolueitten ryhmälle.

Entropiat ryhmien 1 (suuret puolueet) ja 2 (pienet puolueet) sisällä olivat

$$(2.27) \qquad H_W^{(1)} = -\left(\frac{0.239}{0.808} \ln \frac{0.239}{0.808} + \cdots + \frac{0.173}{0.808} \ln \frac{0.173}{0.808} \right) = 1.377$$

ja

$$(2.28) \qquad H_W^{(2)} = -\left(\frac{0.048}{0.192} \ln \frac{0.048}{0.192} + \cdots + \frac{0.016}{0.192} \ln \frac{0.016}{0.192} \right) = 1.553$$

Ryhmien sisäinen entropia kokonaisuudessaan oli siten

$$(2.29) \qquad H_W = 0.808 \cdot 1.377 + 0.192 \cdot 1.553 = 1.411$$

Huomataan, että myös numeroarvojen tasolla pätee dekompositiolaki (2.12), sillä 1.900 = 0.489 + 1.411.

Sisäisten entropioiden (2.27) - (2.29) suhteelliset arvot (suhteellinen arvo laskettuna H_W:n osalta aiemmin esitetyn perusteella niin, että ryhmätason jakauma pidetään kiinnitettynä):

$$(2.30) \qquad \tilde{H}_W^{(1)} = \frac{1.377}{\ln 4} = 0.993$$

(2.31) $$\tilde{H}_W^{(2)} = \frac{1.553}{\ln 5} = 0.965$$

(2.32) $$\tilde{H}_W = \frac{1.411}{0.808 \ln 4 + 0.192 \ln 5} = 0.987$$

Lausekkeiden (2.30) - (2.32) perusteella huomataan, että käyttöönotettu ryhmäjako on tuottanut kaksi sisäisesti (koon suhteen) varsin homogeenista ryhmää, ryhmät ovat lähes tasajakautuneita.

Ryhmittelyn selitysasteeksi muodostuu

(2.33) $$\eta^2 = \frac{1.411}{1.900} = 0.743 ,$$

ts. 74.3 % kokonaisentropiasta voidaan "selittää" ryhmien sisäisellä homogeenisuudella ja ryhmien välisellä heterogeenisuudella (selitysaste olisi 100 % silloin, kun koko kannatus olisi keskittynyt esim. tasajakautuneena suurten puolueiden ryhmään). Aiemmin kuitenkin jo viitattiin niihin ongelmiin, jotka liittyvät η^2-suureeseen. Suhteelliset entropiat (2.24) - (2.26) ja (2.30) - (2.32) muodostavatkin paremman kriteerin muodostettujen ryhmien sisäiselle homogeenisuudelle ja sitä kautta ryhmäjaon onnistuneisuudelle.

2.3. Homogeenisuustestit

Edellä on koko ajan ollut olettamuksena, että jakauma $(p_1, p_2, \ldots, p_n)$ on tarkasti tunnettu, ts. että kyseessä on numeerisen havaintoaineiston tapauksessa kokonaistutkimus (p_i:t populaation suhteellisia frekvenssejä) tai teoreettisessa tarkastelussa

satunnaismuuttujan tunnettu jakauma (p_i:t todennäköisyyksiä). Tässä jaksossa tarkastellaan otoksen perusteella tapahtuvaa ryhmittelyanalyysiä, erityisesti siihen liittyviä testauskysymyksiä. Tarkastelussa on keskeisellä sijalla Kullbackin esittämä informaatiostatistiikka (Kullback 1959: 81-108).

Oletetaan, että käytettävissä on N:n riippumattoman havainnon otos O_N siten, että luokkafrekvenssit ovat $N_1, N_2, \ldots, N_n$, jolloin $N_1 + N_2 + \cdots + N_n = N$. Halutaan testata hypoteesia, että otos on peräisin perusjoukosta, jonka jakauma on

(2.34) $$H_0 : (p) = (p_1, p_2, \ldots, p_n), \sum_{i=1}^{n} p_i = 1 .$$

Vaihtoehtoinen hypoteesi H_1 on tällöin, että otos on peräisin (mistä tahansa) muusta n-luokkaisesta multinomijakaumasta. Kullback on osoittanut (Kullback 1959:112-114), että informaatiostatistiikasta

(2.35) $$I(p; O_N) = \sum_{i=1}^{n} N_i \ln \frac{N_i}{Np_i} .$$

johdetulle testisuureelle $2 I(p; O_N)$ pätee likimain

(2.36) $$2 I(p; O_N) \approx \sum_{i=1}^{n} \frac{(N_i - Np_i)^2}{Np_i} ,$$

jolloin $2 I(p; O_N)$ on likimain $\chi^2(n-1)$ -jakautunut. Informaatiostatistiikka (2.35) on selvästi entropian johdannainen, se on sitä niin ulkonaiselta muodoltaan kuin määritelmälliseltä perustaltaankin. Tämän johdosta tuntuu luonnolliselta odottaa,

että dekompositio-ominaisuuden omaavalla χ^2- jakautuneella testisuureella olisi käyttöä ryhmittelyanalyysissä, joka perustuu niinikään dekompositio-ominaisuuden omaavaan entropia-suureeseen. Näin myös osoittautuu olevan asian laita.

Testattavan nollahypoteesin määrittäminen nyt kyseenä olevassa ryhmittelyanalyysissä on voimakkaasti tapauskohtainen tehtävä. Seuraavassa esitetään kaksi tyypillistä kysymyksenasettelumahdollisuutta.

Ensimmäinen ryhmittelyanalyysin kysymyksenasettelu perustuu Kullbackin (1959: 112-117) esittämään malliin kuitenkin siten yleistettynä, että kahden ryhmän sijasta tarkastellaan yleisesti k ryhmää. Testattava hypoteesi on, että perusjoukon jakauma luokkatasolla on n-arvoinen multinomijakauma (2.34):n mukaisin parametrein. Luokat on edelleen yhdistetty k ryhmäksi, jolloin ryhmätason jakauma on k-arvoinen multinomijakauma (2.5):n mukaisin parametrein. Eli yhdistettynä nollahypoteesi on

$$(2.37)\qquad H_0 : \begin{cases} (p) = (p_1, p_2, \ldots, p_n), \ \sum_{i=1}^{n} p_i = 1 \\ (q) = (q_1, q_2, \ldots, q_k), \text{ missä} \\ q_j = \sum_{r=1}^{g_j} p_{g_0+\ldots+g_{j-1}+r}, \ j=1,2,\ldots,k . \end{cases}$$

Merkintöjen osalta viitataan jakson 2.2 merkintöihin, otokseen liittyvät suureet numeroidaan vastaavalla tavalla. Otoksen ryhmäfrekvenssejä merkitään vielä

$$(2.38)\qquad M_j = \sum_{r=1}^{g_j} N_{g_0+\ldots+g_{j-1}+r}, \ j=1,2,\ldots,k .$$

Saadaan entropia/informaatio-taulu (vrt. varianssitaulu varianssianalyysissä):

Taulukko 2.1. Hypoteesiin (2.37) liittyvä entropia/informaatiotaulu

Entropian/ informaation lähde	Informaatio-statistiikka	Vapaus-asteet
Ryhmien $G_1, G_2, \ldots, G_k$ välinen	$2 \sum_{j=1}^{k} M_j \ln \frac{M_j}{Nq_j}$	$k-1$
Ryhmän G_1 sisäinen	$2 \sum_{r=1}^{g_1} N_r \ln \frac{N_r/M_1}{p_r/q_1}$	g_1-1
. . .	. . .	. . .
Ryhmän G_k sisäinen	$2 \sum_{r=1}^{g_k} N_{g_0+\ldots+g_{k-1}+r} \ln \frac{N_{g_0+\ldots g_{k-1}+r}/M_k}{p_{g_0+\ldots+g_{k-1}+r}/q_k}$	g_k-1
Yhteensä	$2 \sum_{i=1}^{n} N_i \ln \frac{N_i}{Np_i}$	$n-1$

Koko jakaumaa koskeva luokkatason testaus, testisuureena taulukon alimman rivin informaatiostatistiikka (2.36), jonka jakauma H_0:n ollessa tosi on likimain $\chi^2(n-1)$-jakautunut, voidaan nyt halluttaessa hajottaa osiin taulukon 2.1 mukaisesti. Näin erityisesti tapauksissa, joissa H_0 tulee hylätyksi, voidaan löytää poikkeaman lähempi syy. Taulukossa on riveittäin mainittu entropian lähde,

vastaava testisuure (informaatiostatistiikka) ja sen likimain noudattaman χ^2-jakauman vapausasteluku.

Jos ryhmittelyanalyysillä kuitenkin ja nimenomaan tavoitellaan (koon suhteen) homogeenisten ryhmien etsimistä ja osoittamista, ei edellä esitetty kysymyksenasettelun perusmuoto ole useinkaan paras mahdollinen. Homogeenisten ryhmien tapauksessahan ryhmien sisäisten jakaumien tulisi olla tasaiset. Tämä mahdollisuus sisältyy tietysti erikoistapauksena edellä olevaan (p_i:tten määrityksellä). Mutta tällöin H_0:sta tulee käytännön tilanteiden kannalta liian rajoittava: ei ehkä olekaan etukäteen mahdollista hypotetisoida ryhmätodennäköisyyksiä q_j ja näistä sitten luokkatodennäköisyyksiä p_i . Sen sijaan voidaan asettaa väljempi, ryhmittelyanalyysin kannalta kuitenkin riittävän yksityiskohtainen hypoteesi

(2.39) H_0 : ryhmät $G_1, G_2, \ldots, G_k$ sisäisesti tasajakautuneita (otoksesta estimoiduin ryhmätodennäköisyyksin)

eli matemaattisesti

$$(2.40)\qquad H_0 : p_{g_0+\ldots+g_{j-1}+r} = \frac{1}{g_j}\frac{M_j}{N}\,,\quad j=1,\ldots,k, r=1,\ldots,g_j.$$

Kun ryhmätodennäköisyyksien sijasta joudutaan käyttämään niiden estimaatteja, menetetään vapausasteita k-1 kpl. Samalla häviää mahdollisuus ryhmienvälisen entropian testaukseen. Jäljelle jää kuitenkin useasti tärkein selvityksen kohde, homogeenisten ryhmien olemassaolon testaus. Nollahypoteesiin (2.39) liittyvä entropia/informaatiotaulu on nyt

Taulukko 2.2. Hypoteesiin (2.39) liittyvä entropia/informaatiotaulu

Entropian lähde	Informaatiostatistiikka	Vapausasteet
Ryhmien välinen	0	0
Ryhmän G_1 sisäinen	$2\sum_{r=1}^{g_1} N_r \ln \frac{N_r}{M_1/g_1}$	g_1-1
. . .	. . .	. . .
Ryhmän G_k sisäinen	$2\sum_{r=1}^{g_k} N_{g_0+\ldots+g_{k-1}+r} \ln \frac{N_{g_0+\ldots+g_{k-1}+r}}{M_k/g_k}$	g_k-1
Yhteensä	$2\sum_{j=1}^{k}\sum_{r=1}^{g_j} N_{g_0+\ldots+g_{j-1}+r} \ln \frac{N_{g_0+\ldots+g_{j-1}+r}}{M_j/g_j}$	n-k

Esimerkki 2.2. Tarkastellaan esimerkkinä otoksen perusteella tapahtuvasta ryhmittelyanalyysistä seuraavaa (täysin kuviteltua) tilannetta. Oletetaan, että muuttujalla on kaikkiaan kuusi luokkaa, joista kahden ensimmäisen odotetaan esiintyvän huomattavasti yleisemmin kuin neljän jälkimmäisen. Näin muodostuvien kahden ryhmän odotetaan edelleen olevan sisäisesti homogeenisia (tasajakautuneita). Asetetaan nyt testattavaksi kaksi eri nollahypoteesia siten, että ensimmäinen on tyyppiä (2.37) ja toinen tyyppiä (2.39). Ensimmäisessä tapauksessa siis annetaan myös ryhmätodennäköisyydet, jälkimmäisessä testataan pelkästään ryhmien sisäistä tasajakautuneisuutta.

Oletetaan, että otoskoko on N = 200. Otosfrekvenssien oletetaan muodostavan luokka- ja ryhmätasolla taulukon 2.3 mukaiset jakaumat.

Taulukko 2.3. Esimerkkiotoksen luokka- ja ryhmäfrekvenssit

Luokka	E_1	E_2	E_3	E_4	E_5	E_6	Yht.
Frekv. N_i	75	60	25	15	15	10	200
Frekv. M_j	135		65				200
Ryhmä	G_1		G_2				Yht.

Asetetaan nyt tyyppiä (2.37) oleva nollahypoteesi, jossa ryhmätodennäköisyydet q_1 ja q_2 suhtautuvat kuten 3:2. Saadaan

$$(2.41)\quad H_0^{(1)} : \begin{cases} (p) = (0.3,\ 0.3,\ 0.1,\ 0.1,\ 0.1,\ 0.1) \\ (q) = (0.6,\ 0.4) \end{cases}$$

Taulukon 2.3 otokseen ja hypoteesiin (2.41) liittyen saadaan taulukon 2.1 mukaisesti entropia/informaatiotauluksi

Taulukko 2.4. Esimerkin 2.2 entropia/informaatiotaulu (hypoteesina $H_0^{(1)}$).

Entropian lähde	Informaatio-statistiikka	Vapaus-asteet
Ryhmien G_1 ja G_2 välinen	4.81	1
Ryhmän G_1 sisäinen	1.67	1
Ryhmän G_2 sisäinen	7.03	3
Kokonaisentropia	13.51	5

Koska $\chi^2_{0.95}(5) = 11.1 < 13.51$, tulee $H_0^{(1)}$ hylätyksi tasolla $\alpha = 0.05$. Informaatiostatistiikan eri komponentteja tarkastelemalla huomataan kuitenkin, että $\chi^2_{0.95}(1) = 3.84 > 1.67$ ja $\chi^2_{0.95}(3) = 7.81 > 7.03$, ts. ryhmiä G_1 ja G_2 voidaan kuitenkin pitää sisäisesti homogeenisina. Hypoteesin (2.41) hylkääminen ei näin johdukaan siitä, että ryhmät olisivat sisäisesti heterogeeniset (ei-tasajakautuneet), vaan siitä, että ryhmätason jakauma ei ole $H_0^{(1)}$:n mukainen. Sama johtopäätös seuraa informaatiostatistiikan ensimmäisen komponentin tarkastelusta: $\chi^2_{0.95}(1) = 3.84 < 4.81$. Lopullinen johtopäätös ($\alpha = 0.05$) on siten, että luokat E_1 ja E_2 muodostavat homogeenisen ryhmän G_1 ja luokat E_3 - E_6 homogeenisen ryhmän G_2, mutta ryhmätodennäköisyydet q_1 ja q_2 eivät suhtaudu kuten 3 : 2.

Lausekkeen (2.41) mukainen hypoteesi voi tulla kysymykseen lähinnä vain luonnontieteissä, esimerkiksi perinnöllisyystutkimuksessa (6 eri tyyppiä olevia jälkeläisiä, joista kaksi tyyppiä muita yleisempiä; yleisiä ja harvinaisia esiintyy puolestaan keskenään samoissa suhteissa). Yhteiskuntatieteissä sen sijaan lienee tavallisempi tilanne, jossa voidaan kyllä odottaa sisäisesti homogeenisia ryhmiä, mutta ei pystytä etukäteen muodostamaan käsitystä ryhmäosuuksista (esim. kaksi tasavahvaa valtapuoluetta, neljä pienryhmittymää). Tällöin testauksen suorittaminen edellyttää ryhmätodennäköisyyksien estimointia otoksesta, tullaan (2.40):n mukaiseen nollahypoteesiin ja taulukon 2.2. mukaiseen entropia/informaatiotauluun.

Esimerkissä on nyt hypoteesina

$$(2.42)\quad H_0^{(2)} : \begin{cases} p_1 = p_2 = 0.3375 \\ p_3 = p_4 = p_5 = p_6 = 0.08125 \end{cases}$$

ja entropia/informaatiotauluna

Taulukko 2.5. Esimerkin 2.2 entropia/informaatiotaulu (hypoteesina $H_0^{(2)}$.

Entropian lähde	Informaatio-statistiikka	Vapaus-asteet
Ryhmän G_1 sisäinen	1.67	1
Ryhmän G_2 sisäinen	7.03	3
Kokonaisentropia	8.70	4

Nyt on $\chi^2_{0.95}(4) = 9.49 > 8.70$, joten nollahypoteesi jää kokonaisuudessaan voimaan ($\alpha = 0.05$). Samalla tämä merkitsee ryhmien G_1 ja G_2 sisäistä homogeenisuutta. Tulos on tältä tärkeimmältä osaltaan yhdenmukainen edellisen testin tulosten kanssa.

3. RYHMITTELYANALYYSIN SOVELLUTUS: KANSANEDUSTAJAIN VAALIT 1970 - 1983

Jakson 2.2 esimerkissä jo tarkasteltiin entropian komponenttijakoon perustuvan ryhmittelyanalyysin sovellutuksena Suomen eduskuntapuolueiden kannatuksen jakaumaa vuoden 1979 vaaleissa. Tässä pykälässä tarkastelu ulotetaan kattamaan kaikki kansanedustajain vaalit ajanjaksolta 1970 - 1983. Tarkastelu suoritetaan nimenomaan esitetyn ryhmittelyanalyysin sovellutuksena, siis tilastotieteen eikä valtio-opin tai politiikan tutkimuksen näkökulmasta. Analyysi ei sinänsä tuokaan esille varsinaista uutta kannatuksen hajaantumisesta tai keskittymisestä puolue- ja ryhmätasolla, mutta vahvistaa kylläkin yleisen näkemyksen kannatuksen keskittymisestä 1970-luvulla yhä voimakkaammin suurten puolueiden ryhmälle ja tämän suuntauksen päättymisestä, jopa osittaisesta kääntymisestä vuoden 1983 vaaleissa. Lisäksi ryhmittelyanalyysi tuottaa kvantitatiivisen mitan keskittymisen absoluuttiselle ja suhteelliselle voimakkuudelle sekä ryhmien sisäiselle homogeenisuusasteelle.

Vaaleissa mitatut puolue- ja ryhmätason kannatusprosentit on esitetty taulukossa 3.1. Suurten puolueiden ryhmään on luettu kuuluviksi Suomen Sosiaalidemokraattinen Puolue (SDP), Kansallinen Kokoomus (KOK), Keskustapuolue (KESK) ja Suomen Kansan Demokraattinen Liitto (SKDL). Pienten puolueiden ryhmän muodostavat Ruotsalainen Kansanpuolue (RKP), Suomen Maaseudun Puolue (SMP), Kristillinen Liitto (SKL) ja Liberaalinen Kansanpuolue (LKP), jotka ovat saaneet kansanedustajia ajanjakson kaikissa vaaleissa (LKP ei kuitenkaan v. 1983 vaaleissa), sekä yhdistetty luokka MUUT, johon on luettu puolueet, jotka ovat saaneet edustajia

vain joissakin yksittäisissä vaaleissa tai ei lainkaan. Liberaalisen Kansanpuolueen liittymistä Keskustapuolueen jäsenjärjestöksi ennen vuoden 1983 vaaleja ei ole otettu huomioon vaan luokkajaon samana säilymiseksi sitä on v. 1983 vaaleissakin käsitelty erillisenä puolueena. Kannatustiedot perustuvat Mitä-Missä-Milloin -teoksen eri vuosikertoihin (1973: 154-155, 1976: 186, 1980: 146-147), vuoden 1983 vaalien osalta Helsingin Sanomien 23.3.1983 päivättyyn numeroon.

Taulukko 3.1. Puolueiden kannatusprosentit vv. 1970-1983 eduskuntavaaleissa

Puolue/ Ryhmä	Kannatusosuus (%)				
	1970	1972	1975	1979	1983
SDP	23.4	25.8	24.9	23.9	26.7
KOK	18.0	17.6	18.4	21.7	22.1
KESK	17.1	16.4	17.7	17.3	16.6
SKDL	16.6	17.0	18.9	17.9	14.0
"Suuret"	75.1	76.8	79.9	80.8	79.4
RKP	5.7	5.4	4.8	4.5	4.9
SMP	10.5	9.2	3.6	4.6	9.7
SKL	1.1	2.5	3.3	4.8	3.0
LKP	5.9	5.2	4.4	3.7	1.0
MUUT	1.7	0.9	4.0	1.6	2.0
"Pienet"	24.9	23.2	20.1	19.2	20.6
Yhteensä	100.0	100.0	100.0	100.0	100.0

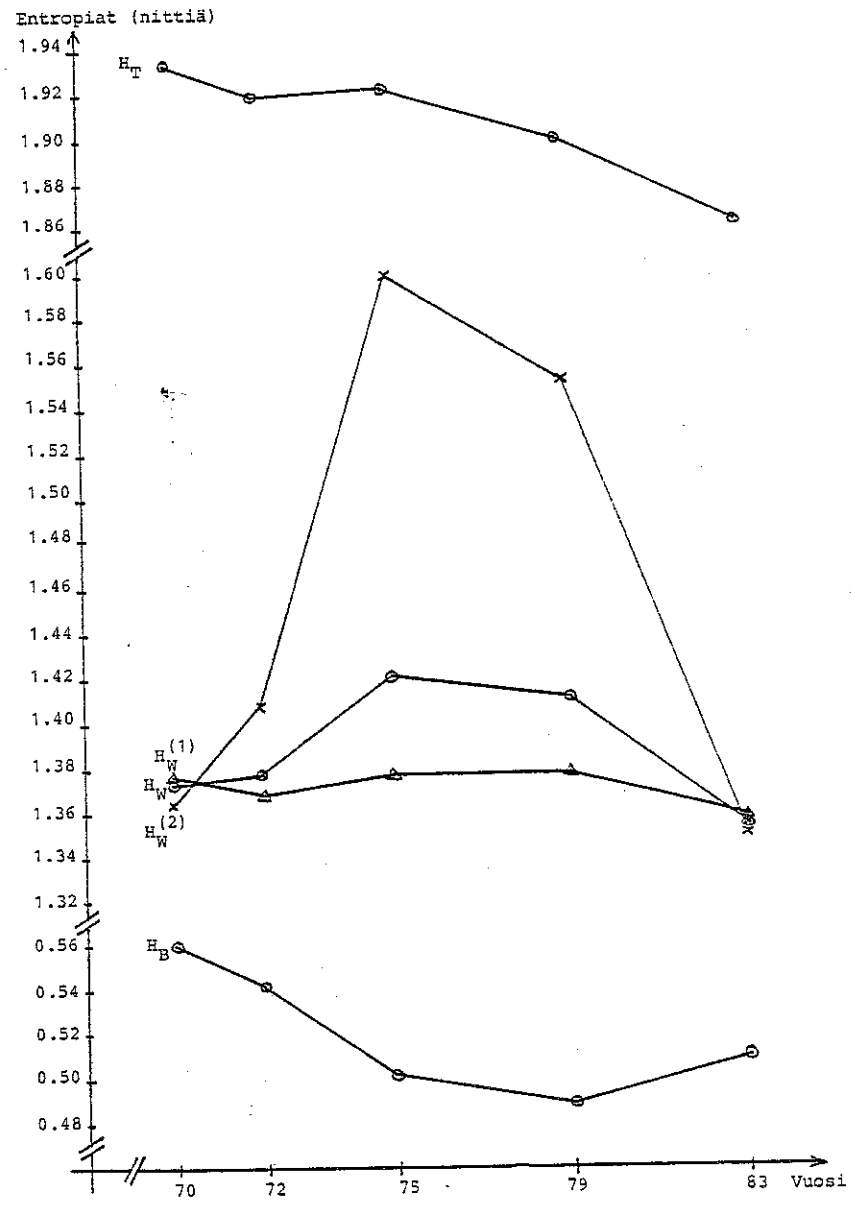
Taulukon 3.1 perusteella on laskettu kuhunkin vaalitulokseen liittyen arvot (sekä absoluuttiset että suhteelliset) suureille

kokonaisentropia H_T, ryhmien "suuret" ja "pienet" välinen entropia H_B, ryhmän "suuret" sisäinen entropia H_{W_1}, ryhmän "pienet" sisäinen entropia H_{W_2}, ryhmien sisäinen kokonaisentropia H_W sekä ryhmittelyn homogeenisuus- 1. selitysaste η^2. Tulokset esitetään taulukossa 3.2 ja kuviossa 3.1.

Taulukko 3.2. Puolueiden kannatusjakaumiin liittyvät entropia-suureet

Tunnusluku	Vaalivuosi				
	1970	1972	1975	1979	1983
H_T	1.934	1.919	1.923	1.900	1.863
$\tilde{H}_T$	0.880	0.873	0.875	0.865	0.848
H_B	0.561	0.542	0.502	0.489	0.509
$\tilde{H}_B$	0.809	0.782	0.724	0.706	0.734
$H_W^{(1)}$	1.376	1.368	1.376	1.377	1.356
$\tilde{H}_W^{(1)}$	0.993	0.987	0.993	0.993	0.978
$H_W^{(2)}$	1.364	1.407	1.600	1.553	1.350
$\tilde{H}_W^{(2)}$	0.848	0.874	0.994	0.965	0.839
H_W	1.373	1.377	1.421	1.411	1.354
$\tilde{H}_W$	0.952	0.958	0.993	0.987	0.945
η^2	0.710	0.718	0.739	0.743	0.727

Tarkastellaan nyt lyhyesti eräitä johtopäätöksiä, jotka ovat tehtävissä taulukon 3.2 ja kuvion 3.1 perusteella. Kokonaisentropian H_T kehitys on ollut koko tarkasteluajanjakson lievästi laskeva. Entropian ominaisuuksien perusteella tämä merkitsee,



Kuvio 3.1. Puoluekannatusjakauman entropian ja sen komponenttien kehitys vv. 1970-83 vaaleissa.

että kannatuksen jakautumisessa on tapahtunut keskittymistä (= loittonemista tasajakautuneisuudesta). Suhteellisen entropian avulla mitattuna tasajakautuneisuusaste on laskenut v. 1970 vaalien 88.0 %:sta v. 1983 vaalien 84.8 %:iin. Koko 1970-luvulla tämä keskittyminen selittyy kokonaisuudessaan, vuosikymmenan alkupuolella jopa ylisuhteisesti, kannatuksen siirtymisellä suurten puolueiden ryhmään: ryhmien välinen entropia H_B on ollut laskeva. Vuoden 1983 vaaleissa ryhmien välisen entropian lasku kuitenkin pysähtyi, H_B:n arvo jopa hieman kohosi. Tältä osin edelleen jatkunut kokonaisentropian lasku selittyykin ryhmien sisäisen entropian H_W pienenemisellä (ryhmien sisäisellä heterogenisoitumisella).

Suurten puolueiden ryhmä on entropialla mitattuna säilynyt koko tarkasteluajanjakson varsin stabiilina. Viimeisissä vaaleissa kannatusosuudet kuitenkin erkanivat selvemmin toisistaan, mikä myös näkyy ryhmän sisäisen entropian $H_W^{(1)}$ laskuna. Ryhmän homogeenisuus on ollut varsin korkealla tasolla, suhteellinen entropia oli 1970-luvulla n. 99 %, v. 1983 vielä 97.8 %.

Suurimmat vaihtelut kannatuksen jakautumisessa ovat sijoittuneet ryhmän "pienet puolueet" sisälle. Täällä vain RKP:n kehitys on ollut vakaata, muilla on suuriakin vaihteluita kannatuksessaan. Erityisen voimakasta kannatuksen vaihtelu on ollut SMP:llä, jonka kannatushuiput sijoittuvat jakson alkuun ja loppuun, aallonpohja taas jakson keskelle. SMP:n kannatuksen vaihtelu on jopa niin voimakasta, että se suorastaan säätelee ryhmän sisäisen entropian kehitystä. Ryhmän tasajakautuneisuusaste on jakson alussa ja lopussa vain vajaa 85 % (v. 1970 84.8 %, v. 1983 83.9 %), kun

taas v. 1975 kannatus oli lähes täysin tasajakautunut ryhmän puolueiden kesken ($H_W^{(2)} = 99.4$ %).

Jaottelun "suuret puolueet" - "pienet puolueet" tuottamien ryhmien sisäistä homogeenisuutta kokonaisuudessaan mittaava ryhmien sisäinen entropia H_W saa muotonsa lähinnä komponentista $H_W^{(2)}$, mutta tasonsa ja vaihtelujensa laajuuden pääasiassa komponentista $H_W^{(1)}$. Kehitys on ollut selvä: 1970-luvulla kuljettiin kohti kahta sisäisesti suhteellisen tasakannatuksista ryhmää, mutta v. 1983 tilanne muuttui ratkaisevasti. Kannatus ryhmien sisällä on epätasaisimmin jakautunutta koko tarkastelukaudella. Tältä osin molemmat ryhmät käyttäytyvät samalla lailla. Kaiken kaikkiaan ryhmäjako on kuitenkin edelleenkin perusteltu. Tasajakautuneisuusaste (suhteellinen sisäinen entropia $\tilde{H}_W$) on kehittynyt v. 1970 vaalien 92.8 %:sta v. 1975 vaalien 97.1 %:n kautta 94.5 %:iin v. 1983. Kehityksen jatkuminen samansuuntaisena voi kuitenkin saada aikaan, että jokin muu ryhmittely, esim. suuret, keskisuuret ja pienet puolueet, johtaa korkeampaan sisäiseen homogeenisuuteen.

Ryhmittelyn selitysaste η^2 käyttäytyy odotetulla tavalla: 1970-luvun kehitys kohti kahta sisäisesti homogenisoituvaa ryhmää näkyy selitysasteen kasvuna 71.0 %:sta 74.3 %:iin, ja v. 1983 tapahtunut käänne laskee selitysasteen 72.7 %:iin. Selitysaste toimii varsin hyvin ryhmittelyä kuvaavana suureena tässä yhteydessä.

Edellä oleva ryhmittelyanalyysi on perustunut puolueiden vaaleissa saamaan kannatukseen. Kyseessä on kokonaistutkimus, joten suureiden arvot ja niissä tapahtuneet muutokset ovat todellisia, jaksossa 2.3 esitettyä päättelykoneistoa ei näin ollen tarvita.

Käytännön politiikan kannalta kannatusosuuksia tärkeämpiä ovat kuitenkin puolueiden saamat kansanedustajapaikat. Suomen vaalijärjestelmästä johtuen näiden jakauma voi melkoisestikin poiketa vaaleissa mitatun kannatuksen jakaumasta, näin erityisesti pienten puolueiden kohdalla. Tästä johtuen edellä suoritettu ryhmittelyanalyysi toistetaan käyttäen perustana kansanedustajien lukumäärien jakaumia. Puolueet ja ryhmät ovat kuten edellä. Jakaumat on esitetty taulukossa 3.3 , analyysin tulokset taulukossa 3.4 ja kuviossa 3.2 .

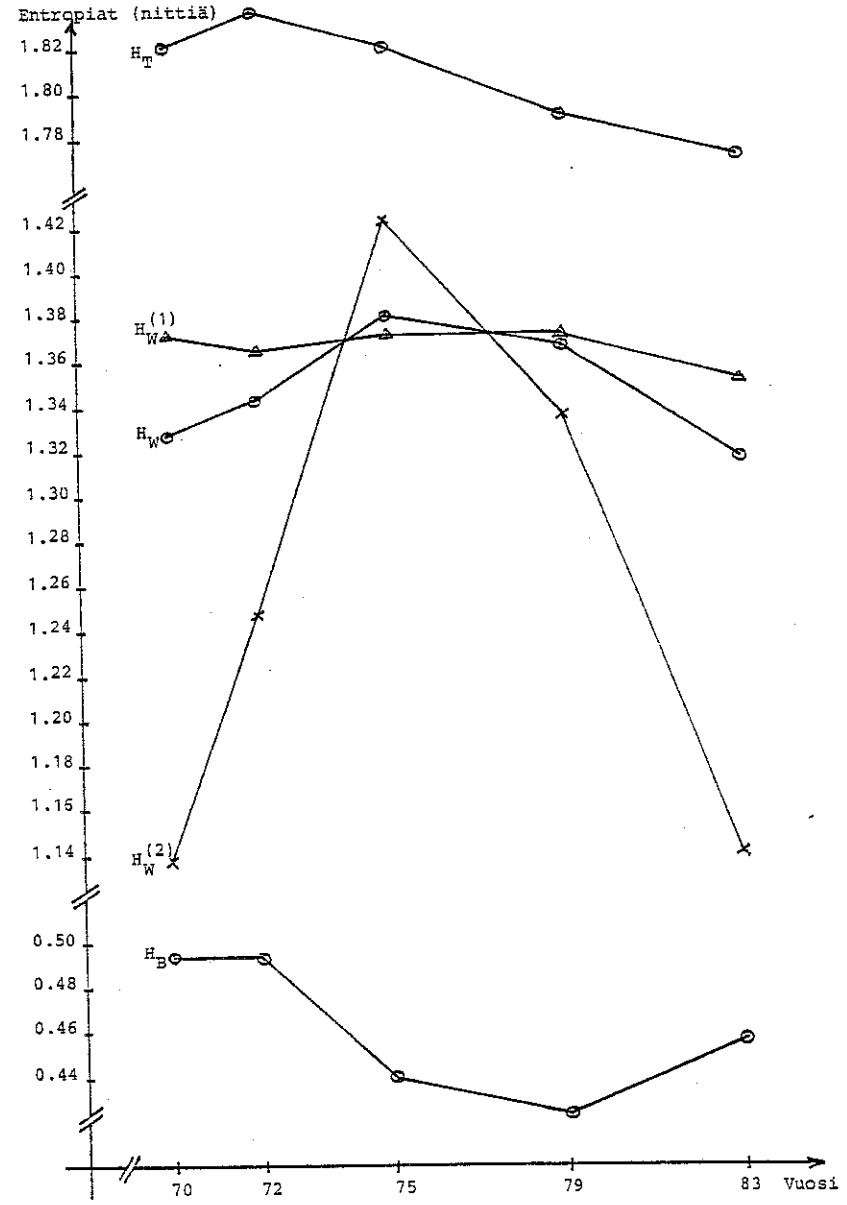
Taulukko 3.3. Kansanedustajien lukumäärät puolueittain vv. 1970-1983 vaaleissa.

Puolue/ Ryhmä	Kansanedustajapaikkoja				
	1970	1972	1975	1979	1983
SDP	52	55	54	52	57
KOK	37	34	35	47	44
KESK	36	35	39	36	38
SKDL	36	37	40	35	27
"Suuret"	161	161	168	170	166
RKP	12	10	10	10	11
SMP	18	18	2	7	17
SKL	1	4	9	9	3
LKP	8	7	9	4	0
MUUT	0	0	2	0	3
"Pienet"	39	39	32	30	34
Yhteensä	200	200	200	200	200

Taulukko 3.4. Kansanedustajien lukumääräjakaumiin liittyvät entropia-suureet.

Tunnus-luku	Vaalivuosi				
	1970	1972	1975	1979	1983
H_T	1.821	1.836	1.820	1.789	1.772
$\tilde{H}_T$	0.829	0.836	0.828	0.814	0.806
H_B	0.493	0.493	0.440	0.423	0.456
$\tilde{H}_B$	0.711	0.711	0.634	0.610	0.658
$H_W^{(1)}$	1.373	1.365	1.372	1.372	1.352
$\tilde{H}_W^{(1)}$	0.990	0.985	0.990	0.990	0.975
$H_W^{(2)}$	1.138	1.248	1.424	1.336	1.140
$\tilde{H}_W^{(2)}$	0.707	0.775	0.885	0.830	0.708
H_W	1.327	1.342	1.380	1.367	1.316
$\tilde{H}_W$	0.928	0.939	0.971	0.963	0.920
η^2	0.729	0.731	0.758	0.764	0.743

Taulukon 3.4 ja kuvion 3.2 tuloksia on syytä kommentoida lähinnä vain siltä osin kuin ne poikkeavat kannatusosuuksien perusteella lasketuista vastaavista arvoista. Kokonaisentropian (puoluetason jakauman entropian) H_T ja ryhmien välisen entropian (ryhmätason jakauman entropian) H_B kehitysurat ovat hyvin samankaltaiset, vain tasoero vallitsee. Paikkojen lukumääräjakaumasta lasketut arvot ovat systemaattisesti pienempiä kuin kannatuksen jakaumasta lasketut. Suhteellisista arvoista laskettuna erot ovat keskimäärin 4.5 %-yksikköä $\tilde{H}_T$:n osalta ja 8.5 %-yksikköä $\tilde{H}_B$:n osalta. Paikkajakauma on siis keskittyneempää kuin



Kuvio 3.2. Kansanedustajien lukumääräjakauman entropian ja sen komponenttien kehitys vv. 1970-83 vaaleissa.

kannatusjakauma, ryhmätasolla vielä enemmän kuin puoluetasolla. Tulos heijastaa vaalijärjestelmän lievähkösti, mutta systemaattisesti suuria puolueita suosivaa paikkojen määräytymismekanismia.

Suurten puolueiden sisäisen jakauman kohdalla ei juurikaan ole merkitystä sillä, onko entropia (= $H_W^{(1)}$) laskettu paikka- vai kannatusjakaumasta: kehitysurat ovat muodoltaan ja tasoltaan lähes identtiset. Suuret puolueet ovat saaneet paikkoja toisiinsa nähden "oikeudenmukaisesti". Pienten puolueiden sisäisen jakauman kohdalla (entropia $H_W^{(2)}$) pätee sama toteamus kuin H_T:n ja H_B:n kohdalla. Paikkajakauman entropian kehitysura on muodoltaan samanlainen kuin kannatusjakauman vastaavan entropian kehitysura, tasoero sen sijaan on nyt melkoinen. Suhteellisissa entropioissa keskimääräinen ero on 12.5 %-yksikköä, sisäinen paikkajakauma pienten puolueiden ryhmässä on selvästi heterogeenisempi kuin vastaava kannatusjakauma. Vaalijärjestelmä on erityisesti tältä osin muuntanut äänestäjien vaaleissa ilmaisemaa tahtoa.

4. YHTEENVETO

Tämän tutkimuksen tavoitteena oli tarkastella entropia-suuretta luokitusasteikon tasoiseen muuttujaan sovellettavan ryhmittelyanalyysin yhteydessä. Tarkastelun lähtökohtana oli entropian ja kvantitatiivisen muuttujan varianssin välinen niin määritelmällinen kuin tulkinnallinenkin analogia. Varianssianalyysin yhteydessä esiintyvä varianssin dekomponointi ryhmien väliseen ja niiden sisäiseen varianssiin antoi aiheen etsiä samanlaista käyttäytymistä myös entropian osalta. Kokonaisentropian jakaantuminen kahteen komponenttiin, tarkasteltavien ryhmien väliseen ja ryhmien sisäiseen entropiaan, olikin löydettävissä ja on esitetty jaksossa 2.2.

Jakso 2.2 sisältää myös entropiakomponentteihin perustuvaa ryhmien homogeenisuustarkastelua. Tarkasteltavan ryhmäjaon homogeenisuusasteelle esitettiin lopuksi tunnusluku, jolla on selvät yhteydet varianssianalyysin yhteydessä käytettyyn selitysasteeseen, joskin tunnusluvun käyttäytymiseen erityisesti arvoalueen päissä liittyy tiettyä epätäsmällisyyttä tai monikäsitteisyyttä. Kokonaisentropian komponenttien, erityisesti niiden suhteellisten arvojen osoitettiin kuitenkin muodostavan hyvän perustan käytetyn ryhmittelyn homogeenisuuden mittaukselle.

Jaksossa 2.3 tarkasteltiin otantaan perustuvaa päättelyä ryhmittelyn tuottamien ryhmien sisäisestä homogeenisuudesta. Testisuureena käytettiin Kullbackin esittämää informaatiostatistiikkaa, jonka käyttö muokattiin ryhmittelyanalyysin tarpeisiin sopivaksi. Jaksossa esiteltiin kaksi testaustilannetta, toisessa hypoteesina

oli luokka- ja ryhmätodennäköisyyksiltään tunnetun jakauman noudattaminen, toisessa luokkatodennäköisyyksien suhteen homogeenisten ryhmien olemassaolo. Erityisesti jälkimmäinen todettiin ryhmittelyanalyysin yhteydessä käyttökelpoiseksi.

Luku 3 muodostaa tutkimuksen sovellutusosan. Siinä tarkasteltiin Suomen eduskuntapuolueiden vaaleissa saaman kannatuksen ja eduskuntaan saamien kansanedustajien lukumäärien kehitystä ajanjaksona 1970 - 1983. Tarkasteltavana ryhmäjakona oli jaottelu "suuret puolueet" - "pienet puolueet". Kokonaisentropiaan ja sen komponentteihin perustuva analyysi toi selvästi esille niin laadullisesti kuin määrällisestikin sekä 1970-luvulla tapahtuneen kehityksen kohti kahta yhä selvemmin toisistaan erottuvaa kokoryhmää että tämän kehityksen pysähtymisen ja muuttumisen osin vastakkaissuuntaiseksikin v. 1983 vaaleissa. Analyysissä heijastui myös vaalijärjestelmän aiheuttama tietynasteinen ero vaaleissa mitatun kannatuksen ja toteutuneiden eduskuntapaikkojen välillä.

SUMMARY

ENTROPY AS A MEASURE OF HOMOGENEITY IN CATEGORICAL GROUPING ANALYSIS

The object of the study is to consider Shannon's entropy (expression (1.1) or (2.1)) as a measure or statistic in categorical variable grouping analysis. The analysis is based on the definitional and interpretational analogy between entropy of a qualitative variable and variance of a quantitative variable (eqs. (2.3) and (2.4); for a more detailed description see Astola and Virtanen (1981: 12-14)). The decomposition property of the total variance in ANOVA leads to expect analogical behaviour for entropy also: if the classes of the variable have been associated to form aggregated groups, the total entropy of the distribution might be decomposed into two components, the entropy between the groups and the entropy within the groups.

The decomposition procedure is presented in Section 2.2. Equation (2.11) shows that the two entropy components, entropy between the groups (H_B given by (2.6)) and entropy within the groups (H_W given by (2.9) and (2.10)) add to the total entropy H_T . The homogeneity of the groups may be analyzed via these entropies: the more H_W contributes to the total entropy, the more homogeneous (i.e. the more uniformity distributed) the groups internally are. The homogeneity of the groups may also be measured by using relative, i.e. between 0 and 1 scaled, entropy quantities (2.19) - (2.21) which eliminate the effect of the number of classes and groups on the entropy quantities.

An index for the homogeneity of the grouping procedure, the degree of homogeneity η^2, is also introduced in Section 2.2 (equation (2.22); cf. R^2 in ANOVA). The section ends with an illustrative example from the finnish representative elections in 1979. In the example the variable categories are the different parliament parties, the groups considered are "the big parties" and "the small parties". The example shows that the elections in 1979 produced two quite homogeneous party groups (with respect to the party size).

Section 2.3 deals with the entropy-based grouping analysis applied to a sample. As a test statistic to test the hypothesis about the homogeneity of the groups there is introduced Kullback's information statistic (2.35) which, being doubled, is under (2.34) asymptotically distributed as χ^2 with n-1 degrees of freedom. For testing homogeneity, two different null hypothesis are considered, viz. (2.37) and (2.40). In (2.37) also the group probabilities of the homogeneous groups are hypothesized whereas in (2.40) the group probabilities have been estimated from the sample. In Tables 2.1 and 2.2 the sources of entropy/information (between the groups, within the groups, total), the corresponding information statistics and degrees of freedom are presented for null hypothesis (2.37) and (2.40), respectively. The section ends with an illustrative example about the homogeneity tests.

Section 3 contains a numerical discussion of the finnish representative elections in 1970 - 1983. The groups considered

are the group of "big parties" and the group of "small parties". The size of the parties is measured both by the (relative) number of votes obtained (Table 3.1) and by the number of representatives elected (Table 3.3). The different entropy quantities are presented in Table 3.2 and Figure 3.1 (based on the distribution of the number of votes) and in Table 3.4 and Figure 3.2 (based on the distribution of the number of representatives). The results show that in 1970's the two groups were distinguished more and more from each others and became internally more homogeneous at the same time. In 1983, however, this process of development ceased. We can further see that the results based on the distribution of Table 3.1 and Table 3.3 are qualitatively similar but quantitavely slightly different, especially for the group "small parties". This difference is due to the finnish electoral system.

LÄHDELUETTELO

Astola, Jaakko - Virtanen, Ilkka (1981). Entropy correlation coefficient, a measure of statistical dependence for categorized data. Tutkimusraportti, Lappeenrannan teknillinen korkeakoulu.

Astola, Jaakko - Virtanen, Ilkka (1982). A measure of overall statistical dependence based on the entropy concept, Vaasan korkeakoulun julkaisuja, Discussion papers 44.

Astola, Jaakko - Virtanen, Ilkka (1983). A measure of overall statistical dependence based on the entropy concept, Vaasan korkeakoulun julkaisuja. Tutkimuksia No 91.

Georgescu-Roegen, Nicholas (1967). Analytical Economics. Cambridge, Massachusetts: Harvard University Press.

Helsingin Sanomat 23.3.1983.

Horowitz, Ann R. - Horowitz, Ira (1976). The real and illusory virtues of entropy-based measures for business and economic analysis. Decision Sciences 7, 121-136.

Kettunen, Pertti (1973). Laskentatoimen informaation tarkkuudesta, arvosta ja määrästä. Luentomoniste. Jyväskylän yliopisto, taloustieteen laitos.

Kullback, Salomon (1959). Information theory and statistics. New York: John Wiley & Sons.

Latosaari, Erkki (1983). Entropiakäsite taloudellisissa ja tilastollisissa sovellutuksissa. Julkaisematon ekonomitutkinnon tutkielma. Vaasan korkeakoulu.

Mitä-Missä-Milloin 1973 (1972). Helsinki: Otava.

Mitä-Missä-Milloin 1976 (1975). Helsinki: Otava.

Mitä-Missä-Milloin 1980 (1979). Helsinki: Otava.

Pihlajarinne, Eero (1979). Informaatiokertoimet ja niiden käyttö ryhmittelyanalyysissä. Jyväskylän yliopiston julkaisuja.

Preuss, L.G. (1980). A class of statistics based on the information concept. Communications in statistics, theory and methods 9:15, 1563-1586.

Shannon, C.E. (1948). A mathematical theory of communication. Bell System Technical Journal 27, 379-423, 623-656.

Theil, Henri (1969). On the use of information theory concepts in the analysis of financial statements. Management Science 15:9, 459-480.

Walker, M.C. - Stowe, I.D. - Moriarty, S. (1979). Decomposition analysis of financial statements. Journal of Business Finance and Accounting 6:2, 173-186.

Wilson, A.G. (1970a). Entropy in urban and regional modelling. London: Pion Limited.

Wilson, A.G. (1970b). The use of concept of entropy in system modelling. Operational Research Quarterly 21:2, 247-265.