Tilastotieteilijä tarvitsee matematiikkaa – entä matemaatikko tilastotiedettä?

Seppo Laaksonen
Matematiikan ja tilastotieteen laitos
Helsingin yliopisto

Palasin yliopistomaailmaan vuonna 2002 pidemmän poissaolon jälkeen. Opetuskokemusta on nyt kertynyt sekä yleisiltä että erityisiltä kursseilta sekä ohjauksesta. Yllättävää on ollut huomata, ettei tilastotieteen asema ole kohentunut yliopistossa vaikka työelämässä jatkuvasti olen havainnut alan osaajien puutteen. Myös olen hämmästellyt sitä, että perusylioppilas tietää tilastotieteestä edelleen vähän ja monet pääaineopiskelijatkin ovat tulleet alalle sattumalta, ilman intohimoa. Väistämättä tämän täytyy johtua kouluopetuksen luonteesta.

PISA-tulosten mukaan Suomen yläasteikäiset koululaiset pärjäävät yhä mainiosti matemaattis-tilastollisessa lukutaidossa, mitä nimeä taidan tosin ainoana käyttää. Monet PISA-tehtäväthän ovat tilastollisia, jopa hieman todennäköisyyksiinkin viittaavia eli ei sitä matematiikkaa mitä kunnon matemaatikot rakastavat.

Oltakoonpa terminologiasta mitä mieltä tahansa, niin olen huolestunut ylioppilassukupolvien matemaattisesta osaamisesta. Minulle kerrotun mukaan yksi kolmannes ylioppilaskokelaista ei osallistu minkäänlaiseen matemaattiseen tenttiin ja suorittajista osa ei kuulemma käytännössä osaa juuri mitään. Tämä näkyy yliopistojen ja varmaan myös ammattikorkeakoulujen kursseilla, joissa muun muassa tilastotieteen perusopinnot ovat pakollisia monille. Jopa peruslaskutoimitusten ja suuruussuhteiden ymmärtämisessä on suuria vaikeuksia, saati sitten että vaikkapa integrointi ja derivointi onnistuisivat.

Jotain siis olisi syytä tehdä. Yksi perusvaatimukseni olisi asettaa ainakin jokin matemaattis-tilastollinen alue pakolliseksi ylioppilaille. Huolella toki pitäisi miettiä mikä tai mitkä olisivat sopivia alueita. Toinen ehdotukseni on opetuksen motivoinnin parantaminen siten, että matematiikan sovellus ja siis hyöty tulisivat entistä paremmin esille. Tämän kirjoituksen jatkossa esitän muutamia ajatuksia ja myös konkreettisia esimerkkejä tältä näkökulmalta.

Esitelkää matematiikan käsitteiden yhteyksiä käytäntöön

En tunne tarkasti lukion matematiikan oppisisältöjä. Tilastotieteellistä otetta siellä on joka tapauksessa liian vähän. Mutta kaikille matematiikan oppiaineksille on helppo löytää käytännön kytkentöjä. Opettajille tutuimpia lienevät fysiikan tai kemian kytkennät. Toivottavasti niitä tuodaan esille.

Tilastollisia kytkentöjä on varmasti myös kaikkialla. Esimerkiksi integrointi johtaa empiirisen aineiston piirissä summaamiseen, jossa yhteydessä käytetään summamerkkiä. Tämä näyttää olevan ylioppilaille kummajainen. Logaritmia ei enää nykysukupolvi hahmota suh-

teellisen mittaamisen upeana välineenä. Itsehän tämän opin laskutikun kautta. Tästä syystä tilastollisessa grafiikassa esiintyy jatkuvasti huonoja asteikkoja, siis absoluuttisia suhteellisten sijasta. Vastaavasti harhaudutaan eksponentin ymmärtämättömyyden takia toiseen suuntaan. Polynomit ovat myös paljon käytettyjä, osin logaritmien ja eksponenttien rinnalla. Tämän artikkelin loppuosa keskittyy polynomeihin pyrkien havainnollistamaan näiden hyötykäyttöä.

Polynomit, niiden derivointi ja ääriarvot

Polynomit ovat kivoja funktioita. Yksinkertaisin vaihtoehto on puhdas vaakasuora mitä jossain yksinkertaisessa tilanteessa käytetään tilastotieteessä, jolloin se merkitsee esimerkiksi keskiarvoa tai mediaania. Vielä yleisempi polynomi on suora, josta tilastotieteessä käytetään nimitystä lineaarinen. Jos se derivoidaan, saadaan vakio eli suoran kulmakerroin. Useamman asteisille polynomeille ei tilastotieteessä tietääkseni ole erityisiä nimiä. Seuraavaksi esitän kolme esimerkkiä, joissa vähintään esiintyy toisen ja kolmannen asteen polynomeja. Nämä ovat aika yleisiä tilastotieteessä ja sen sovellustieteissä kuten talous- ja sosiaalitieteissä.

Esimerkki 1: Ikäonnellisuus

Onnellisuuden tutkimus on yleistynyt erityisesti psykologiassa ja taloustieteissä. Kun aihetta tutkitaan empiirisesti, tarvitaan tilastollinen aineisto. Tavallisesti aineisto koostuu ihmisille esitetyistä kysymyksistä. Tässä esitettävä tulos perustuu Euroopan yhteiskuntatutkimuksen (Europeansocialsurvey.com) 15 vuotta täyttäneistä suomalaisista kerättyyn haastatteluaineistoon vuosilta 2002-2007.

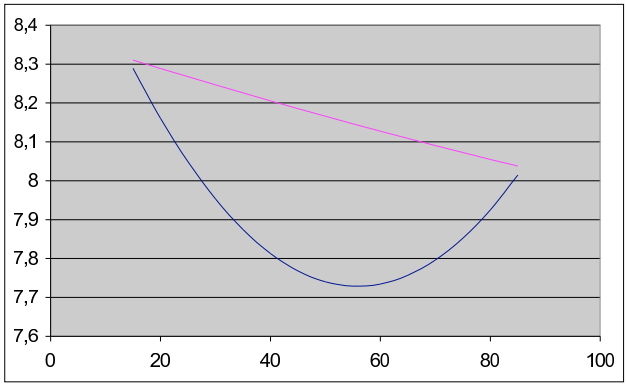
Onnellisuuden taustatekijöistä suuri kiinnostus on kohdistunut ikään. Taloustieteilijät ovat havainneet, että monesti ikäonnellisuus noudattaa ns. U-käyrää eli onnellisuus on nuorena korkea, laskee sitten keski-ikään mennessä jolloin alkaa taas nousta. Yksilöaineistosta tutkittuna tämä käyrä tarkoittaa paraabelia. Tilastotieteilijä tutkii asiaa asettamalla malliin kaksi selittäjää, iän ja sen neliön. Tämän jälkeen hän estimoi sen ja katsoo tuloksista, onko väitteellä perää.

Selitin ihmisen kokemaa onnellisuutta (asteikko 0-10) tilastollisella mallilla ikä ja sen neliö selittäjinä, kummallekin sukupuolelle erikseen. Estimointitulokset ovat seuraavassa:

$$\begin{aligned}&\text{onnellisuus(naiset)}\\ &\quad = 0{,}000006476\,\text{ikä}^2 - 0{,}0045424\,\text{ikä} + 8{,}3775\end{aligned}$$

$$\begin{aligned}&\text{onnellisuus(miehet)}\\ &\quad = 0{,}000336286\,\text{ikä}^2 - 0{,}037554\,\text{ikä} + 8{,}7772\end{aligned}$$

Kuvio 1 havainnollistaa tilannetta graafisesti. Tästä näemme että naisten ja miesten onnellisuus on melko sama nuorena ja vanhana mutta miesten onnellisuus laskee selvästi nuoruuden jälkeen. Lukion matematiikan opeilla on helppo laskea minimi-iät, ensin derivoimalla ja sitten ratkaisemalla nollakohdat; tee se. Tulos miehille on 55,8 vuotta ja naisille 350 vuotta



Kuvio 1. Onnellisuus paraabelilla estimoituna naisille (ylempi käyrä) ja miehille (alempi).

Kuviosta ja minimistä on helppo nähdä, että miesten käyrä on jossain määrin U:n muotoinen mutta naisten ei, vaikka siis estimointi antaa ylöspäin aukeavan paraabelin. Ei ole kuitenkaan järkeä ajatella naisten käyrän olevan U-mainen. Miksi? Käyrä on mieluummin lähes lineaarinen.

Tässä esimerkissä esitin vain matemaattisen näköisen puolen, samoin tapahtuu esimerkissä 2. En siis keskustele paraabeliin liittyvää epävarmuutta mitä siihen tietystikin liittyy. Käyrällä on siis tosiasiassa tietty luottamusväli, samoin kuin minimiarvoissa.

Esimerkki 2: Ikä ja palkka

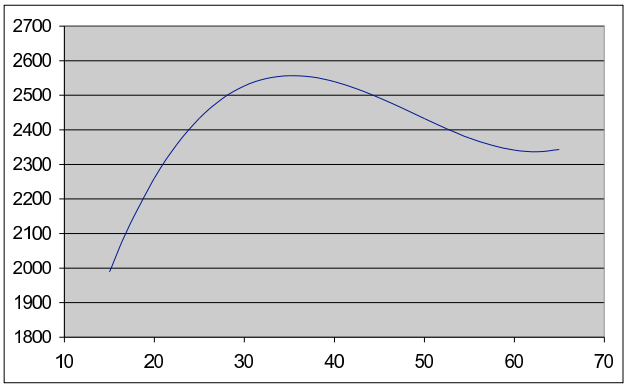
Toinen esimerkki on eräänlainen laajennus edelliselle. Nyt käytössä on kolme selittäjää, ikä, sen neliö ja sen kuutio. Siten muodostuva funktio on kolmatta astetta. Selitettävänä on palkansaajan kuukausipalkka eräässä aineistossa.

Ihan samalla periaatteella kuin esimerkissä 1 estimoin yhtälön:

$$\text{palkka} = 0{,}0225\,\text{ikä}^3 - 3{,}292\,\text{ikä}^2 + 148{,}6$$

Vastaavasti tein Kuvion 2. Kaikki kolme muuttujaa ovat merkitseviä, mikä antaa edellytyksen uskoa että palkkakäyrässä on sekä minimi että maksimi. Nämä voidaan ratkaista derivoimalla ja sitten ratkaisemalla nollakohdat. Huippukohta saavutetaan tällä aineistolla varsin nuorena eli 35,4 vuoden iässä. Tämän jälkeen

palkka laskee mutta alkaa nousta juuri ennen tavallista eläkeikää eli 62,3 -vuotiaana (selitykseni on se, että korkeapalkkaiset jatkavat työelämässä pidempään). Tarkista tulosten oikeellisuus.



Kuvio 2. Palkka kolmannen asteen käyrän funktiolla estimoituna.

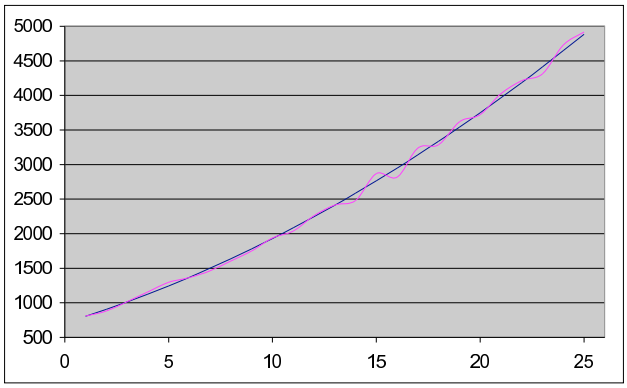
Matalimmillaan palkka on toki työelämään siirryttäessä, mikä on tässä asetettu 15 vuoden kohdalle mistä saakka havaintoja oli aineistossa, joskin vähän. Vanhimmat palkansaajat tässä ovat 64-vuotiaita. Matemaattisesti käyrä voidaan piirtää näiden ikien ulkopuolelle, mutta tilastollisesti ei ole niin syytä tehdä. Onnellisuuskuviossa asetin käyrän välille 15-85 -vuotiaat, vaikka vanhimmat vastaajat olivat 99-vuotiaita. Jos haluat, jatka käyrää tänne asti.

Esimerkki 3: Aikasarja

Tämä esimerkki ei ole todellinen mutta tähtää havainnollistamaan todellisuutta. Muodostin 25 havaintoyksikköä, jotka merkitty ajankohtina t. Toiseksi tein teknisen muuttujan x joka saa arvoja 15:sta 39:een yhden yksikön välein. Tässä on siis yksinkertainen aritmeettinen sarja.

Varsinaisen aikasarjamuuttujan muodostin toisen asteen polynomilla $3x^2+8x+10$. Huomaa että tämän ensimmäinen derivaatta $= 6x+8$ ja toinen $= 6$. Taulukossa 1 tätä aikasarjaa merkitsen symbolilla y. Se on siis funktiomuotoinen ja ajattelen sen tässä olevan suurin piirtein estimoitu oikeista havaintoarvoista yr. Oikeat havaintoarvot eivät koskaan noudata mitään funktiomuotoa mutta voivat olla lähellä sellaista. Tutkijan jatkotyö on helppoa, jos löytää aineistossa funktiomaisen yhteyden. Tässä esimerkissä tilanne on hoidettu niin, että yhteys on varsin hyvä. Katso itse tätä Kuviosta 3.

Havaitsemme kuviosta ehkä selkeämmin kuin taulukosta, että aikasarja kasvaa kiihtyvästi. Kuvio muistuttaa eksponentiaalista kasvua, sillä paraabelilla on sopivilla parametriarvoilla samanlaisia ominaisuuksia. Voit itse tehdä oman kokeesi eksponenttifunktiomuotoa käyttämällä.



Kuvio 3. Todellinen aikasarja ja sen funktiomuoto.

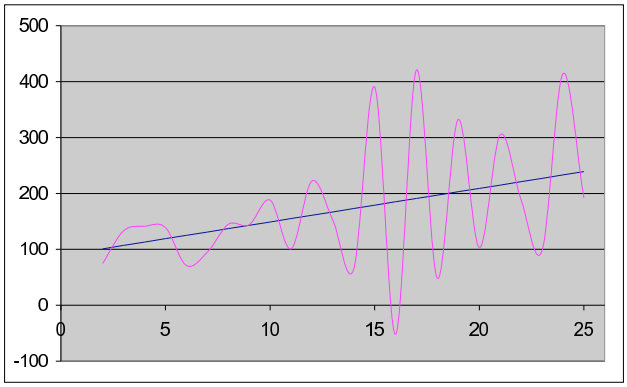
Aikasarjaa voi tutkia monin tavoin. Tässä tutkitaan muutosta mikä esimerkissä tarkoittaa kasvua. Analoginen mutta päinvastainen tilanne koskee vähenemistä.

Laskin kummallekin aikasarjalle aritmeettiset muutokset eli differenssit (asiaa voisi tutkia myös suhteellisesti):

diff1y = y:n arvon muutos
ajankohdasta t ajankohtaan $t+1$

diff1yr = yr:n arvon muutos
ajankohdasta t ajankohtaan $t+1$

Taulukosta näemme, että y:n differenssisarja on nyt aritmeettinen, arvot kasvavat edellisestä aina 6:lla, mikä on ensimmäisen derivaattafunktion kulmakerroin ja toisen derivaattafunktion vakiotermi. Todellinen aikasarjani ei ole näin kaunis, vaan vaihtelut ovat suurehkoja, keskiarvokin jää noin 5:een. Teoria ei siis täysin istu todellisuuteen mikä on ymmärrettävää. Kuvio 4 havainnollistaa tätä eroa.



Kuvio 4. Ensimmäiset differenssit todelliselle ja teoreettiselle sarjalle.

Teoreettinen sarja on lineaarinen ja sen kulmakerroin on siis 6. Tämä viiva asettuu kuitenkin hyvin todellisten havaintoarvojen keskelle. Todellisista estimoitu kulmakerroin on 6,1 eli lähellä teoreettista todellisuutta.

Aikasarja-analyysissä on tapana ottaa toiset differenssit eli ensimmäisten differenssien differenssit. Taulukossa nämä on merkitty seuraavasti:

$\text{diff2}y = \text{diff1}y$:n arvon muutos
ajankohdasta t ajankohtaan $t+1$

$\text{diff2}yr = \text{diff1}yr$:n arvon muutos
ajankohdasta t ajankohtaan $t+1$

Havaitsemme että teoreettisen sarjan arvot ovat vakioita eli siis toisen derivaatan arvoja. Tämä osoittaa että aikasarjan y kasvu ei ole kiihtyvää vaan on aivan tasainen. Todellisessa sarjassa ei nytkään havaita yhtä kaunista asetelmaa. Muutosten muutokset vaihtelevat huomattavasti mutta mitään selvää trendiä niistä ei havaita. Tämä siis myös osoittaa ettei kasvu ole kiihtyvää. Jos haluat, voit piirtää tästä osasta vastaavan kuvion kuin edellä.

Tässä esimerkissäni käytin funktiomaista aikasarjaa jotta derivoinnin ja differenssioinnin yhteys näkyy hyvin. Kokeile muilla funktiomuodoilla vastaavaa myös. Käytännössä ei siis löydy hyvää funktiomuotoa millä tilanteen näkisi yksinkertaisesti. Differenssioinnin sen sijaan voi aina tehdä. Jos toisen differenssin arvoissa havaitset ylöspäin menevää trendiä, kasvu on kiihtyvää; jos se näyttäisi menevän alaspäin, kasvu on hidastuvaa (kuten taloustieteilijät äskettäin uskoivat Suomessa tapahtuvan). Vähenemisen puolella voidaan käyttää vastaavia termejä. Esimerkiksi hidastuva väheneminen tai alaspäinmeno jossakin asiassa merkitsee monelle jo positiivista signaalia.

t	x	y	yr	diff1y	diff1yr	diff2y	diff2yr
1	15	805	804				
2	16	906	878	101	74		
3	17	1013	1011	107	133	6	59
4	18	1126	1151	113	140	6	7
5	19	1245	1289	119	138	6	-2
6	20	1370	1360	125	71	6	-67
7	21	1501	1455	131	95	6	24
8	22	1638	1599	137	144	6	49
9	23	1781	1742	143	143	6	-1
10	24	1930	1929	149	187	6	44
11	25	2085	2030	155	101	6	-86
12	26	2246	2250	161	220	6	119
13	27	2413	2402	167	152	6	-68
14	28	2586	2466	173	64	6	-88
15	29	2765	2854	179	388	6	324
16	30	2950	2803	185	-51	6	-439
17	31	3141	3221	191	418	6	469
18	32	3338	3269	197	48	6	-370
19	33	3541	3600	203	331	6	283
20	34	3750	3702	209	102	6	-229
21	35	3965	4005	215	303	6	201
22	36	4186	4193	221	188	6	-115
23	37	4413	4290	227	97	6	-91
24	38	4646	4702	233	412	6	315
25	39	4885	4893	239	191	6	-221

Taulukko 1. Aikasarjani aineisto ja sen muunnokset.