Tilastollinen otanta (Kaius Sinnemäki)

Transkriptio

Tässä videossa käsittelen otantaa ja esittelen täällä Excelissä kaksi erityyppistä otantaa, eli yksinkertainen satunnaisotanta ja lisäksi ositettu otanta. Olen laatinut tänne tämmöisen havaintomatriisin, jossa on koulujen nimiä Vantaan lähiöittäin, muutamia lähiöitä siellä on, ja sitten joitakin luokkia sinne on valittu. Koululuokkia, 1A 1B niin poispäin.

Jokaiselle havaintoyksikölle on annettu myös oma tunnuksensa. Huomaa, että tämä havaintoaineisto on niin sanotun havaintomatriisin muodossa, eli sarakkeissa ovat muuttujat ja sitten riveillä ovat havaintoyksiköt. Ensimmäinen rivi sisältää näitten muuttujien nimet ja siitä eteenpäin on sitten havaintoyksiköittäin nuo muuttujia koskevat havaintoarvot.

Kun toimitaan Excelissä ja aineistoa on enemmän kuin yhden näytöllisen verran, eli kun sitä joutuu vierittämään, niin on kätevää kiinnittää tuo ensimmäinen rivi. Tällä tavoin. Eli sitten kun vieritetään, niin otsikkorivit säilyvät tuolla paikoillaan.

Eli otetaan ensin yksinkertainen satunnaisotanta. Tämä Excel, jota käytän tässä, on englanninkielinen, ja Excelit voivat muilla kielillä poiketa hieman tästä, mitä selvitän nyt, mutta selitän kyllä nämä funktiot sekä englanniksi että suomeksi. Meillä on tässä, kun katotaan näitä tunnuksia, ne kulkee juoksevassa numerossa eli meillä on yhteensä 56 tapausta tai 56 havaintoa tässä aineistossa. Me voidaan käyttää siis tämmöistä funktiota Excelissä kuin RANDBETWEEN. Eli valitaan se täältä listasta, joka aukeaa.

RANDBETWEEN-funktio saa kaksi arvoa, eli ensin sille annetaan numerojonon ensimmäinen eli pienin luku ja sitten numerojonon suurin luku. Eli meillä esimerkiksi tää tunnus kulkee luvusta 1 lukuun 56, niin annetaan ne luvut tähän. Eli pienin on 1, käytetään puolipistettä, otetaan 56, sulkeet kiinni, painetaan enteriä. Ja nyt funktio RANDBETWEEN valitsi satunnaisesti väliltä 1:stä 56:een yhden luvun ja tuo luku nyt oli 45. Sit me voitais mennä täältä alemmas, etsiä 45 ja sieltä löytyy Kaivokselan luokka 1A ja voitais sitten laittaa tänne muistiin. Nyt huomasitte varmaan, että heti kun täytin jonkun muun solun arvon, niin tämä RANDBETWEEN antoi heti uuden satunnaisluvun. Tämän vuoksi on tärkeää toimia tän RANDBETWEEN-funktion kanssa sillä tavalla, että se voidaan kopioida toiseen soluun eli olen tämän solun kohdalla painanut Ctrl-C ja sitten siirryn edelliseen soluun, painan Ctrl-V, ja sitten se antaa tämmöisen valikon, että miten tämä solu ja sen arvot nyt sitten tänne

liitetään. Niin täältä kohdasta Paste Values otan tämän, missä näkyy vain 1-2-3 eli nyt se liittää pelkät arvot. Eli kun tarkastellaan tätä solua, niin nähdään, että sen sisältö on tuo funktio, mutta tämä viereinen solu mihin me liitettiin, niin siellä on pelkästään nyt sitten tuo se funktion tuottama arvo. Eli me voitais sitten ottaa uudestaan 32 tuolta: Tikkurila 2C, ja siitä tulisi meille ensimmäinen satunnaisesti valittu arvo.

Nyt näistä funktion nimistä, suomeksi tämä RANDBETWEEN-komento on muistaakseni =satunnaisluku.väliltä(). Jos käytät jotain muunkielistä Exceliä, niin kannattaa etsiä hakukoneesta RANDBETWEEN ja sitten se kieli, jolla haluat sen komennon. Mutta suomeksi muistaakseni se on =satunnaisluku.väliltä(), mutta voit vielä itse tarkistaa sen.

No, jos me haluttais valita tästä aineistosta sanotaanko vaikka viisi tapausta satunnaisesti, niin me voidaan tehdä tämä seuraavasti. Eli mennään tänne soluun, jossa tämä RANDBETWEEN-funktio on, painetaan Ctrl-C eli kopioidaan, ja sitten maalataan viisi solua ja painetaan Ctrl-V eli liitä.

Jos me halutaan, niin me voidaan sitten, tai oikeastaan on hyvä tähän vaikka viereen ottaa uusiksi ne ja liittää vain nuo arvot, mitä (jaha, nyt se ei tietysti, nyt se tietysti sitten antoi kaikki samaa lukua, otetaas uudestaan, katotaas mitä nyt tapahtuu, liitetään… noin, nyt se antoi kaikki erikseen sieltä eli ei yhtä ja samaa lukua). No, nyt me voidaan sitten etsiä nämä eli 14, katotaan Myyrmäki 2C, sitten 18 Myyrmäki 4A, sitten 31 (se oli tuolla väärä) eli 31 ja 33, Tikkurila. Noin ja sitten 42 vielä, katotaan mikä sieltä tulee. Eli sieltä saatiin Tikkurila 6A. Eli näin me ollaan satunnaisesti valittu viisi tapausta tästä aineistosta. Eli tämä on yksinkertainen satunnaisotos.

No tässä huomataan nyt, että meillä on ollut neljä koulua täällä. Meillä on Hakunila, Myyrmäki, Tikkurila ja Kaivoksela. Mutta kun me tehtiin yksinkertainen satunnaisotos, niin me ei saatu yhtään tapausta Hakunilasta eikä yhtään tapausta Kaivokselasta, vaan me saatiin satunnaisesti valittuna jotenkin ainoastaan tapauksia Tikkurilasta ja Myyrmäestä.

Voidaan ajatella, että tässä on sellanen mahdollisuus, että tämä kuitenkin jollain lailla vinouttaa meidän otosta. Voitais ajatella, että olisi hyvä saada tapauksia myös Hakunilasta ja Myyrmäestä. No tähän on olemassa erilaisia vaihtoehtoja, miten tämä tehdään. Yksi tapa on esimerkiksi se, että jos me haluttais vain yksi luokka Hakunilasta, yksi luokka Myyrmäestä, yksi Tikkurilasta ja yksi Kaivokselasta, niin me voitais käyttää RANDBETWEEN-funktiota sillä tavalla, että annetaan vain se väli näitten lukujen väli, mikä koskee sitä kyseistä koulua. Eli esimerkiksi Hakunilassa nämä tunnukset kulkevat 1:stä

8:aan, niin me voidaan silloin tehdä näin eli Hakunilasta valittaisiin tässä tapauksessa 8 eli Hakunila 4B. Sitten Myyrmäen kohdalla kopioidaan tuo RANDBETWEEN, mutta tuota alle. Luvut kulkevat 9:stä 26:een. Mennään tänne funktioriville, laitetaan sinne 9 puolipiste 26 ja kopioidaan se tuohon viereen pelkästään. Se arvo, jonka se tuottaa sieltä eli 10. Otetaan Myyrmäki 1C. Ja näin me voitais edetä ja ottaa jokaisesta koulusta yksi luokka satunnaisesti.

Tämä kuitenkin on vähän hidasta siinä mielessä varsinkin sellaisessa tapauksessa, että halutaan jokaisesta ositteesta eli jokaisesta koulusta useampi tapaus. Niin näytän nyt sen, miten tämmönen ositettu otanta tehtäis, kun halutaan useampi tapaus jokaisesta ositteesta. Monesti kun otetaan ositettu otos, niin silloin halutaan myös ottaa niitä havaintoja ositteista suhteessa siihen, että miten paljon niitä havaintoja ylipäätänsä on kussakin ositteessa. Eli esimerkiksi meillä on täällä kahdeksan luokkaa Hakunilasta, mutta meillä onkin kun maalataan kaikki nämä Myyrmäen kohdat, niin nähdään meillä on 18 tapausta Myyrmäeltä. Eli jossain mielessä olisi mielekästä valita useampia luokkia Myyrmäestä ja harvempia Hakunilasta, joten näytän tämmöisen suhteellisen ositetun otannan ottamisen seuraavaksi.

Edetään niin, että mennään tänne välilehdelle Insert, luodaan PivotTable eli painetaan PivotTable, ja sitten valitaan täältä aineistosta niin, että nyt mä painan Shift eli koululuokka ja sit meen nuolella alaspäin. Valitaan tuo koko aineisto tuosta noin, ja sit painetaan OK. Sit meille tuli tämmöinen PivotTable Fields -ikkuna tänne oikealle. Otetaan sieltä Luokka, koska me halutaan nyt laskea, kuinka monta luokkaa… anteeksi siis Koulu täältä, eli halutaan laskea, kuinka monta luokkaa koulussa on. Laitetaan se tänne kohtaan Rows, vedetään se sinne. Ja sit tuo Luokka, voidaan se vetää tuonne kohtaan Values.

Niin nyt Excel automaattisesti loi meille tämmöisen taulukon, jossa luokkien määrät koulussa laskettiin automaattisesti. Voidaan sulkea tuosta tuo kenttä. Eli nyt me tiedetään näitten luokkien määrät kouluissa. Seuraavaksi voidaan laskea näitten suhteelliset osuudet. Eli jos ajatellaan, että 56 on se luokkien määrä tässä koko aineistossa, niin lasketaan kuinka monta prosenttia 8 on tuosta 56:sta. Eli painetaan yhtäsuuruusmerkkiä, valitaan 8, sitten Shift-7 saadaan kauttaviiva, ja sitten valitaan tuolta tuo 56, sulkeet kiinni… siellä on joku virhe. Valitaan siis uusiksi. Eli yhtäsuuruusmerkki, valitaan tuo 8, sit jaetaan 56 sieltä valitaan ja sit painetaan enteriä. Noin. Ja sama tehdään näille kaikille. Eli yhtäsuuruusmerkki valitaan tuosta vierestä, jaetaan, ja sit tuolla Totalin määrillä. Sit voidaan laittaa yhtäsuuruusmerkki ja täällä englanninkielisessä Excelissä SUM eli lasken kaikki nuo yhteen, suomenkielisessä muistaakseni lukee SUMMA. Eli sieltä tuli yhteensä 1 niin kuin pitikin tulla noista suhteellisista osuuksista.

Sit mä valitsen kaikki nuo suhteelliset solut, painan hiiren oikeata näppäintä ja painan Format Cells. Muutetaan nuo prosenteiksi eli täältä Prosentit ja otetaan sinne vaikka yksi desimaali ja painetaan OK:ta. Nyt me saadaan näkyviin, miten monta… miten suuri osuus näistä luokkien kokonaismäärästä esiintyy missäkin koulussa.

No nyt me voitais sanoa, että me halutaan joku tietynsuuruinen otoskoko. Eli sanotaan, että me halutaan vaikka 10 luokkaa valita satunnaisesti, niin tehdään näin: aloitetaan uusi funktio yhtäläisyysmerkillä, ja sitten valitaan tuolta toi suhteellinen osuus ja kerrotaan se 10:llä. Eipäs tullutkaan., meidän täytyy varmaan tehdä uudestaan näin, että valitaan tuolta 8, jaetaan se 56:lla ja kerrotaan sitten se tuolla otoskoolla. Nyt sinne saatiin jotakin vähän sen näköistä mitä pitääkin. Eli otetaan täältä nämä suhteelliset osuudet vähän niin kuin uudestaan ja kerrotaan sitten tuolla otoskoolla. Ja sen sijaan, että kirjoitettaisiin jokaisen funktiorivin perään otoskoko 10, niin valitaan tuolta tuo solu erikseen sen takia, että jos me muutetaan sitten sitä otoskokoa, mikä me halutaan, niin se automaattisesti päivittää nuo, että kuinka monta luokkaa mistäkin koulusta otetaan. Ja sitten vielä lasketaan yhteen täältä näiden solujen arvot ja muutetaan nämä myös kokonaisluvuiksi. Otetaan Format Cells, Number nolla desimaalia ja OK.

Eli nyt nähdään, että jos meidän otoskoko on 10 ja me halutaan tehdä ositettu otanta ja nimenomaan suhteellinen ositettu otanta, niin meidän pitäis (ja meidän otoskoko on 10) niin silloin meidän pitäisi valita yksi luokka Hakunilasta, kaksi luokkaa Kaivokselasta ja kolme luokkaa sekä Myyrmäestä että Tikkurilasta. Tällöin me saadaan (no, tässä itse asiassa nähdään, niitä on yhdeksän, mutta kun sitten tuolta yhdistetään niin tulee 10), jos me halutaan vaikka 20 luokkaa, muutetaan tänne otoskoon viereiseen soluun 20, niin se laskee automaattisesti nyt sitten, että miten monta luokkaa mistäkin koulusta sitten valitaan.

Jos me haluttaisiin vaikka sanotaanko nyt että kahdeksan luokkaa ja havaitaan, että Tikkurilasta pitäisi ottaa kolme luokkaa. Sanotaan siis, että otetaan Tikkurila ja sieltä pitää ottaa kolme luokkaa. Käytetään sitten tuota komentoa RANDBETWEEN, kopioidaan se täältä, tuodaan se tänne viereen ja annetaan sille sitten… katotaan missä Tikkurila on. Se alkaa tapauksesta 27, menee tapaukseen 44 asti. Eli 27 on alin arvo ja 44:ään asti. Noin, sitten kopioidaan tänne, liitetään ja valitaan, että liitetään… ei nyt tietysti sitten onnistunut noin. Tota, tuodaan sitten viereen tuosta noin, ja sit katotaan, että tuosta vaikka [epäselvää] halutaan ne, noin. Kiinteiksi arvoiksi sinne. Eli 35, 43 ja 32. Eli tuolta 32 on 2C. Tonne noin, sit 35 3C. Ja sit 43 vielä 6B.

Eli näin me tehdään ositettua, suhteellista ositettua otantaa. Tässä on tämmöistä manuaalista työtä aika paljon, mutta otanta kannattaa tehdä huolellisesti, siihen kuluu joka tapauksessa jonkin verran aikaa aina. Mutta tässä näitä periaatteita nyt, miten yksinkertainen satunnaisotanta tehdään ja miten suhteellinen ositettu otanta tehdään. Tässä esimerkissä käytin nyt kouluja ja niissä olevia luokkia, mutta aivan yhtä hyvin tätä voi soveltaa korpustutkimukseen vaikkapa niin, että koulun sijaan täällä olisi yksittäinen puhuja tai vaikkapa jokin tietty tekstilaji. Luokka voisi olla ihan hyvin yksittäinen puhuja, ja täällä voisi olla ehkä muitakin ositteita sitten. Tai jos puhutaan kielitypologiasta, niin koulu voisi olla vaikkapa kieliperhe, ja luokka voisi olla vaikkapa sitten kieli. Niin sillä lailla saadaan sitten jokaisesta kieliperheestä valittua tietty määrä kieliä.

Tämä tällä erää otannasta.