Kendallin τ ja Spearmanin ρ

W. Hämäläinen

April 25, 2013

1 Ongelman asettelu

Annettuna kaksi numeerista–ordinaalista muuttujaa X ja Y, jotka saavat datassa arvot $(x_1, y_1), \ldots, (x_n, y_n)$. Mikä on X:n ja Y:n korrelaatio annetulla järjestyskorrelaation mitalla?

2 Goodmanin ja Kruskalin Gamma (G tai Γ)

Huom! Goodmannin ja Kruskalin Gamma-kerrointa kutsutaan joskus myös τ:ksi. (Mitat ovat hyvin samantapaisia.)

Perusidea: Tutkitaan kaikki mahdolliset parit (x_i, y_i) ja (x_j, y_j) $i = 1, \ldots, n$ ja $j = 1, \ldots, n$, $i \neq j$. (Tällaisia pareja on yhteensä $\frac{n(n-1)}{2}$ kappaletta.) Vertailu voi tuottaa kolmenlaisia tuloksia:

- Jos ($x_i < x_j$ ja $y_i < y_j$) tai ($x_i > x_j$ ja $y_i > y_j$), ovat parit yhdenmukaisia (concordant).
- Jos ($x_i < x_j$ ja $y_i > y_j$) tai ($x_i > x_j$ ja $y_i < y_j$), ovat parit ristiriitaisia (discordant).
- Jos $x_i = x_j$ tai $y_i = y_j$, **ei pari ole kumpaakaan. Tämä on siis hieman epäintuitiivista: vaikka meillä olisi $x_i = x_j$ ja $y_i = y_j$, kuten funktionaalisessa riippuvuudessa pitää olla, ei paria pidetä yhteensopivana korrelaatiota laskiessa!** Tällaisia pareja kutsutaan kytketyiksi (niiden välillä on kytkös tai sidos, "tie"; datarivit ovat siis (ainakin) kyseisen muuttuja-arvon osalta duplikaatteja).

Näistä laskemme kaksi muuttujaa: yhdenmukaisten parien lukumäärä N_c ja ristiriitaisten parien lukumäärä N_d. Näistä saadaan gammakerroin:

$$G = \frac{N_c - N_d}{N_c + N_d}.$$

Gamman perusmuodossa ei toisiinsa kytkettyjä pareja oteta lainkaan huomioon. Tämä kuitenkin pienentää datan määrää ja tekee mitasta siten epävakaamman. Siksi kytkösten käsittelyyn on kehitetty erilaisia strategioita.

3 Kendallin τ

Kendallin τ muistuttaa hyvin paljon Gammaa. Perusmuodossaan τ on

$$\tau = \frac{N_c - N_d}{\frac{1}{2}n(n-1)}.$$

Mikäli muuttujien välillä ei ollut kytköksiä, on siis $\tau = G$. τ:sta on kuitenkin olemassa erilaisia variaatioita kytkösten käsittelyyn.

τ_a ei huomioi kytköksiä millään tavalla (eli nyt $\tau_a < G$).

τ_b puolestaan on

$$\tau_b = \frac{N_c - N_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}},$$

missä $n_0 = \frac{n(n-1)}{2}$, $n_1 = \frac{1}{2}\sum_i t_i(t_i - 1)$, $n_2 = \frac{1}{2}\sum_i u_i(u_i - 1)$ ja $t_i = |\{x_i\}|$ (moniko alkio saa X:n arvon x_i) ja $u_i = |\{y_i\}|$ (moniko alkio saa Y:n arvon y_i). Ts. n_1 kertoo, monessako parivertailussa on X:n suhteen sidoksia ja n_2 monessako Y:n suhteen.

4 Spearmanin ρ

Spearmanin ρ on käytännössä Pearsonin korrelaatiokerroin muuttujien järjestysnumeroille. Kerroin lasketaan seuraavasti: ensin korvataan kaikki x_i:t ja y_i:t niiden suuruusjärjestysnumeroilla (joko nousevassa tai laskevassa suuruusjärjestyksessä, pääasia että X:n ja Y:n kohdalla noudatetaan samaa järjestystä). Mikäli muuttuja-arvojen välillä on kytköksiä, annetaan kaikille kytketyille arvoille niille kuuluvien järjestysnumerojen keskiarvo. Siis jos esim. $x_m = x_{m+1} = \ldots = x_{m+s}$ ja niille kuuluvat järjestysnumerot olisivat $m, m+1, \ldots, m+s$, annetaan jokaiselle "järjestysnumeroksi" $\frac{1}{s+1}\sum_{j=0}^{s}(m+j)$. Kun kaikki lukuarvot on muunnettu järjestysnumeroiksi, lasketaan vain Pearsonin korrelaatiokerroin. (Tosin näyttäisi, että jopa Spearmanista on erilaisia variaatioita kytkösten käsittelyyn...)

5 Milloin käyttää mitäkin?

Wikipedia suosittelee käyttämään Gammaa ja Kendallin τ:ta vain jos molemmat muuttujat ovat ordinaalisia. Spearmanin ρ:ta taas sanotaan voitavan soveltaa niin jatkuville kuin diskreeteille numeerisille muuttujille, mukaan lukien ordinaalimuuttujat (kun esitetty numeroina).

Lähteessä [Uni] ordinaalimuuttujat jaetaan kahteen ryhmään. Ns. jatkuvat ordinaalimuuttujat tarkoittavat sellaisia ordinaalimuuttujia, joilla on paljon mahdollisia arvoja ja vähän kytköksiä (duplikaattiarvoja). Tämän tyyppiselle datalle suositellaan Spearmanin ρ:ta. Ns. romautettu ordinaalimuuttuja (collapsed ordinal varaible) taas tarkoittaa sellaista ordinaalimuuttujaa, jolla on vähän erilaisia arvoja (korkeintaan 5–6) ja sen tähden myös paljon kytköksiä (duplikaattiarvoja). Tämän tyyppiselle datalle suositellaan Gamma-kerrointa.

Lähde [XHea09] puolestaan pitää τ:ta ja ρ:ta Pearsonin korrelaatiota sopivampina jopa jatkuvalle datalle, mikäli datassa on paljon ulkopuolisia tai riippuvuus on selvästi epälineaarinen. Kirjoittajat mainitsevat että τ:ta pidetään yleensä ρ:ta robustimpana. Oman analyysinsä perusteella he suosittelevat Spearmanin ρ:ta ja gammaa, jos dataa on hyvin vähän (alle 20 riviä) mutta muuten Kendallin τ:ta pidetään näitä parempana. Lisäksi kerrotaan, että ρ on τ:ta selvästi huonompi, mikäli oikea korrelaatio hyvin vahva. Samoin jos datassa on paljon kohinaa (tai ulkopuolisia), on τ suositeltavampi, ainakin jos oikea korrelaatio vähintään kohtuullinen (ei lähellä 0.aa). Joissain tilanteissa paras mitta saattaa kuitenkin olla kirjoittajien esittelemä τ:n ja ρ:n yhdistelmä. ("Oikea korrelaatio" viittaa tässä ilmeisesti Pearsonin korrelaatioon koko populaatiossa, josta data on vain jotenkin vääristynyt otos.)

References

[Uni] Harding University. Ch 14 ordinal measures of correlation: Spearman's rho and gamma. `http://www.harding.edu/sbreezeel/460\%20Files/Statbook/CHAPTER14.pdf`.

[XHea09] W. Xu, Y. Hou, and et al. Comparison of spearmans rho and kendalls tau in normal and contaminated normal models. 2009. Submitted manuscript.