

Ei-parametriset ja robustit menetelmät, kevät 2015

Harjoitus 1

1. Olkoon d_1, \dots, d_n on satunnaisotos jatkuvasta symmetrisestä jakaumasta, jonka mediaani on Δ . Nollahypoteesin $H_0 : \Delta = 0$ testaamiseen voidaan käyttää luennolla esiteltyä Wilcoxonin parittaisten otosten testisuureta

$$W = \#\{w_i > 0\} = \sum_{i \leq j} I\left(\frac{d_i + d_j}{2} > 0\right),$$

jossa w_1, \dots, w_N ovat kaikki mahdolliset pareittaiset keskiarvot $(d_i + d_j)/2$. Määritellään testisuure

$$W_1 = \sum_{i=1}^n R_i^+ I(d_i > 0),$$

jossa R_i^+ on $|d_i|$:n järjestysluku joukossa $D = \{|d_1|, \dots, |d_n|\}$.

- a) Laske testisuureiden W ja W_1 arvot aineistolle $\{-4, -3, 0, 2, 5\}$.
- b) Laske testisuureiden W ja W_1 arvot aineistolle $\{-4, -3, 0, 3, 5\}$.
- c) Oletetaan, että aineistossa ei esiinny toistensa vastalukuja. Osoita, että tällöin $W = W_1$.

Vihje: Olkoon d_{i_1}, \dots, d_{i_p} aineiston positiiviset alkiot. Havainnon $|d_{i_j}|$ järjestysluku joukossa D saadaan laskemalla niiden alkioiden d_i lukumäärä, joiden etäisyys origosta on korkeintaan $|d_{i_j}|$.

2. Tarkastellaan aineistoa, joka on julkaistu kuuluisassa artikkelissa Student (1908). The Probable Error of a Mean. *Biometrika* **6**, 1–25.

Halutaan tutkia eroaako uunikuivatuilla maissin siemenillä saatu sadon määrä tavallisilla siemenillä saadusta sadosta. Kokeessa jaettiin 11 viljelyaluetta kah-tia ja toiseen puoleen kylvettiin tavallisia siemeniä ja toiseen uunikuivattuja sie-meniä. Kokeen tuloksena saatiin taulukon 1 mukaiset satomäärät. Havaintoai-neisto löytyy myöskin tiedostosta `kilndry.data` (muuttujat: `REG`="tavallinen", `KILN`="uunikuivattu"). Laske sekä käsin että tietokoneen avulla t-testin, merk-kitestin ja Wilcoxonin testin testisuureet ja tutki kaksisuuntaisen testin p-arvon avulla eroavatko satomäärät tilastollisesti merkitsevästi. Laske myöskin käsittelyn vaikutuksen keskimääräiselle erolle Δ vastaavat estimaatit ja (liki-main) 95% luottamusvälit. Tehtävässä voi käyttää hyväksi tehtävän 1 tuloksia, luentomonisteen R-esimerkkejä ja valmiita R-funktioita:

t-testi: R-funktio `t.test`

Merkkitesti: R-funktio `SIGN.test` löytyy pakkauksesta `BSDA`

Wilcoxonin testi: R-funktio `wilcox.test`

Taulukko 1: Viljelyalueelta saatu maissin määrä (paunaa/eekkeri) käytettäessä tavallisia ja uunikuivattuja siemeniä

Viljelyalue	Siemenlaatu	
	Tavallinen	Uunikuivattu
1	1903	2009
2	1935	1915
3	1910	2011
4	2496	2463
5	2108	2180
6	1961	1925
7	2060	2122
8	1444	1482
9	1612	1542
10	1316	1443
11	1511	1535

3. Kappaleessa 2.1.5 tappiofunktion D_{II} minimoimisen yhteydessä saatiin tulos

$$\sum_{i=1}^n \text{sign}(d_i - \Delta) = 2 \left(\#\{d_i > \Delta\} - \frac{n}{2} \right).$$

Edellistä yhtäsuuruutta johdettaessa oletettiin, että $d_i \neq \Delta$ kaikilla $i = 1, \dots, n$. Johda vastaava tulos yleisessä tapauksessa (eli d_i saa olla myös yhtäsuuri kuin Δ). Piirrä sen jälkeen funktion $g(\Delta) = \sum_{i=1}^n \text{sign}(d_i - \Delta)$ kuvaaja tapauksissa $n = 4$ ja $n = 5$. Tarkastele kohtia, joissa funktio saa arvon nolla tai vaihtaa merkkiä.