To-utvalg permutasjonstest

TMA4240/TMA4245 Statistikk

Håkon Tjelmeland
Institutt for matematiske fag
Norges teknisk-naturvitenskapelige universitet

To-utvalg permutasjonstest

- Anta
  - vi er gitt $n+m$ pasienter (ikke tilfeldig trukket)
  - to behandlinger: tradsjonell behandling og ny behandling
  - ønsker å vite om ny behandligen er bedre enn den tradisjonelle
  - når en pasient har fått behandling kan vi observere en størrelse x
    - anta at lav verdi av x er ønskelig
    - f. eks: x er rekonvalenstid
    - f. eks: x er mål på hvor mye bivirkninger man opplever
  - formulerer problemet som en hypotesetest

- Hypoteser: H_0 : effekten av behandlingene er like mot H_1 : ny behandling er bedre

- Stokastisk forsøk:
  - trekker tilfeldig n pasienter fra de $n+m$ pasientene tilgjengelige
  - de n pasientene vi trakk ut får tradisjonell behandling
  - de resterende m pasientene får ny behandling
  - observerer størrelsen x for hver av de $n+m$ pasientene

- Stokastiske variabler:
  - $X_1, \ldots, X_n$: observerte verdier for pasienter som får tradisjonell behandling
  - $X_{n+1}, X_{n+2}, \ldots, X_{n+m}$: observerte verdier for pasientene som får ny behandling

- Merk:
  - $X_1, X_2, \ldots, X_{n+m}$ er stokastiske fordi vi randomiserer behandlingen
  - en indikator for om den nye behandlingen er bedre enn den gamle kan være

$$U = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{m}\sum_{i=n+1}^{n+m} X_i$$

  - stor verdi for U indikerer at ny behandlingen er best

- Kan vi bruke U som testobservator?

Hvilken fordeling har U når H_0 er sann?

- Husk:
  - H_0 : effekten av behandlingene er like
  - definisjonen av U

$$U = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{m}\sum_{i=n+1}^{n+m} X_i$$

  - U er stokastisk fordi vi har trukket tilfeldig hvem som får hvilken behandling

- Hvis H_0 sann:
  - vi observerer de samme verdiene uansett type behandling
  - U er likevel stokastisk

- Lekeeksempel: Anta $n = 3$, $m = 2$, altså $n + m = 5$ pasienter
  - antall mulige inndelinger av 5 pasienter i to grupper, 3 i den første og 2 i den andre

$$\binom{5}{3,2} = \frac{5!}{3! \cdot 2!} = 10$$

pasient id	verdi	1	2	3	4	5	6	7	8	9	10
i	12.03	t	t	t	t	t	t	n	n	n	n
ii	9.94	t	t	t	n	n	n	t	t	t	n
iii	10.24	t	n	n	t	t	n	t	t	n	t
iv	11.49	n	t	n	t	n	t	t	n	t	t
v	8.82	n	n	t	n	t	t	n	t	t	t
verdi for u		0.58	1.62	-0.60	1.87	-0.35	0.69	0.13	-2.09	-1.05	-0.80

- Husk: stor verdi for U indikerer at ny behandling er best
- p-verdien blir: $p = P(U \geq 1.62 | H_0 \text{ sann}) = \frac{2}{10} = 0.2$

To-utvalg permutasjonstest

- Anta $n+m$ pasienter totalt
  - n pasienter får tradisjonell behandling
  - m pasienter får ny behandling

- Antall mulige inndelinger av $n+m$ pasienter i to grupper, n i første og m i andre

$$\binom{n+m}{n,\, m} = \frac{(n+m)!}{n! \cdot m!}$$

- La $u^\star$ være observert verdi for testobservator
- La $u_i, i = 1, 2, \ldots, \binom{n+m}{n,\, m}$ være verdi for testobservator for hver av de mulige inndelingene i to grupper
- Anta stor verdi av testobservator indikerer at H_1 er korrekt
- p-verdien blir

$$p = \frac{1}{\binom{n+m}{n,\, m}} \sum_{i=1}^{\binom{n+m}{n,\, m}} \mathbb{I}(u_i \geq u^\star)$$

- MEN: Hva gjør vi hvis $n+m$ er stor?
  - eksempel: $n = 20,\; m = 23$

$$\binom{43}{20,\, 23} = \frac{43!}{20! \cdot 23!} = 960\,566\,918\,220$$

- Kan estimere p-verdien ved stokastisk simulering!

To-utvalg permutasjonstest og stokastisk simulering

- Anta $n + m$ pasienter totalt
  - n pasienter får tradisjonell behandling
  - m pasienter får ny behandling
- Antall mulige inndelinger av $n + m$ pasienter i to grupper, n i første og m i andre

$$\binom{n+m}{n, m} = \frac{(n+m)!}{n! \cdot m!}$$

- La $u^\star$ være observert verdi for testobservator
- Trekk tilfeldig s gruppeinndelinger (uavhengig av hverandre)
  - la u_i være verdi av testobservator ved å bruke i-te gruppeinndeling
- Vi har da $s + 1$ verdier av testobservatoren: $u^\star, u_1, u_2, \ldots, u_s$
- Anta stor verdi av testobservator indikerer at H_1 er korrekt
- Estimert verdi for p-verdien blir

$$\hat{p} = \frac{1}{s+1}\left(1 + \sum_{i=1}^{s} \mathbb{I}(u_i \geq u^\star)\right)$$

Oppsummering

- ⋆ Har sett på to-utvalg permutasjonstest
  - antar ingenting om type fordeling av observasjoner
  - det er (vår) tilfeldige inndeling i grupper som gir tilfeldighet
  - testobservatoren er en diskret stokastisk variabel
- ⋆ Hvis antall mulige gruppeinndelinger er veldig stor
  - ikke overkommelig å beregne eksakt p-verdi
  - kan estimere p-verdien ved stokastisk simulering