Tilastollinen päättely II, syksy 2015 – kevät 2016
Harjoitus 9 (2. ja 4. 2. 2016)

Tehtävät 1–3 liittyvät tyhjentävyyteen. Monisteen luku 4.

1. Tarkastellaan harjoituksen 8 tehtävän 2 mallia, jossa $Y_1, \ldots, Y_n$ olivat riippumattomia ja noudattivat kukin jatkuvaa jakaumaa, jonka tiheysfunktio oli $f(y;\theta) = \theta y^{\theta-1}$, kun $0 \le y \le 1$. Johda faktorointikriteerin avulla jokin yksiulotteinen tyhjentävä tunnusluku parametrille θ.

Ratkaisu 1

Harjoituksen 8 tehtävästä 2 tiedetään, että mallin tiheysfunktio on

$$f(\mathbf{y};\theta) = \theta^n \prod_{i=1}^{n} y_i^{\theta-1},$$

joka riippuu aineistosta vain tunnusluvun $\prod_{i=1}^{n} y_i$ kautta, joka on siis tyhjentävä tunnusluku.

2. (Monisteen teht. 4.5.) Olkoon $f_{\mathbf{Y}}(\mathbf{y};\theta)$ tilastollinen malli, jonka parametrilla on yksikäsitteinen suurimman uskottavuuden estimaatti $\hat{\theta}$. Oletetaan, että $\mathbf{T} = \mathbf{t}(\mathbf{Y})$ on tyhjentävä tunnusluku. Päättele, että $\hat{\theta}$ riippuu aineistosta $\mathbf{y}$ vain tunnusluvun $\mathbf{t}(\mathbf{y})$ välityksellä.

Ratkaisu 2

Koska $\mathbf{t}(\mathbf{y})$ on tyhjentävä tunnusluku, tiheysfunktio voidaan kirjoittaa muodossa

$$f_{\mathbf{Y}}(\mathbf{y};\theta) = h(\mathbf{y})g(\mathbf{t}(\mathbf{y});\theta).$$

Kertomalla parametrista θ riippumattomat tekijät pois, saadaan uskottavuusfunktioksi

$$L(\theta;\mathbf{y}) = g(\mathbf{t}(\mathbf{y});\theta),$$

joka riippu aineistosta vain tunnusluvun $\mathbf{t}(\mathbf{y})$ välityksellä. Uskottavuusfunktiolla on yksikäsitteinen suurimman uskottavuuden estimaatti $\hat{\theta}$, eli maksimiarvo pisteessä $\hat{\theta}$, joten myös $\hat{\theta}$ riippuu aineistosta vain tunnusluvun $\mathbf{t}(\mathbf{y})$ välityksellä.

3. (Monisteen teht. 4.6.) Olkoon $Y_1, \ldots, Y_n$ riippumaton otos jakaumasta, jonka ptf/tf on $f(y;\theta)$. Päättele esim. faktorointikriteerin avulla, että järjestystunnusluku $\big(Y_{(1)}, \ldots, Y_{(n)}\big)$ on aina tyhjentävä.

Ratkaisu 3

Tehtävän tilastollinen malli on riippumattomuuden nojalla muotoa

$$f_{\mathbf{Y}}(\mathbf{y};\theta) = \prod_{i=1}^{n} f(y_i;\theta).$$

Merkitään $y_{(i)}$:llä i:nneksi suurinta havaintoa. Yhteistiheysfunktio tai yhteispistetodennäköisyysfunktio on reaaliarvoinen funktio, joten kertolaskun järjestys voidaan vaihtaa, joten tilastollinen malli voidaan kirjoittaa yhtäpitävästi muodossa

$$f_{\mathbf{Y}}(\mathbf{y};\theta) = \prod_{i=1}^{n} f(y_{(i)};\theta) = g((y_{(1)},\ldots,y_{(n)});\theta).$$

Nähdään, että faktorointikriteerin nojalla järjestystunnusluku on tyhjentävä.

Tehtävät 4 ja 5 liittyvät testiteorian perusteisiin. Monisteen jaksot 5.1–5.3.

4. Seuraavassa on lueteltu sattumanvaraisessa järjestyksessä joitakin tilastollisen testin suorittamiseen liittyviä työvaiheita. Opiskeltuasi itsellesi uudet käsitteet luentomonisteesta, pohdi, missä järjestyksessä vaiheet tulisi suorittaa:

a) testisuureen valinta
b) aineistonkeruu (tai satunnaiskokeen suorittaminen)
c) p-arvon laskeminen
d) tilastollisen mallin määrittely
e) merkitsevyystason asettaminen
f) nollahypoteesin määrittely
g) päätös nollahypoteesin hyväksymisestä/hylkäämisestä

Onko järjestys yksikäsitteisesti määrätty? Mitkä vaiheet väistämättä riippuvat toisistaan ja mitkä taas eivät saisi riippua toisistaan? Ovatko kaikki vaiheet täysin välttämättömiä?

Ratkaisu 4

Testaamisen vaiheiden järjestys ei ole yksiselitteisesti määrätty, mutta seuraavassa yksi mahdollinen järjestys:

d) tilastollisen mallin määrittely
f) nollahypoteesin määrittely
a) testisuureen valinta
e) merkitsevyystason asettaminen
b) aineistonkeruu (tai satunnaiskokeen suorittaminen)
c) p-arvon laskeminen
g) päätös nollahypoteesin hyväksymisestä/hylkäämisestä

Näistä kohdat e) ja g) eli merkitsevyystason asettaminen ja päätös nollahypoteesin hyväksymisestä eli hylkäämisestä liittyvät päätösteoreettiseen lähestymistapaan (ks. monisteen luku 5.3.2) eli ne ovat välttämättömiä vain silloin kun on tehtävä päätös joko hylätä tai hyväksyä nollahypoteesi. Toisaalta taas kohta c), eli p-arvon laskeminen, ei ole välttämätöntä, jos ollaan kiinnostuneita vain siitä, jääkö nollahypoteesi voimaan. Voidaan nimittäin vain laskea testisuureen arvo, ja katsoa onko se testin kriittisellä alueella valitulla merkitsevyystasolla, ja sen perusteella nollahypoteesi joko voidaan hylätä tai se jää voimaan.

Monisteen mukaan tilastollinen malli määritellään ennen nollahypoteesiä, mutta onhan mahdollista että tutkijalla on mielessä nollahypoteesi jo etukäteen, ja hän suunnittelee koeasetelman (mallin) vasta sen jälkeen.

Nollahypoteesin, käytetyn testisuureen ja merkitsevyystason pitäisi teoriassa olla riippumattomia aineiston erityispiirteistä, eli tässä mielessä ne tulevat ennen aineistonkeruuta. Käytännössä kuitenkin usein saatetaan ensin kerätä aineisto ja sitten vasta suunnittelemaan testejä. Tällöin kannattaa kuitenkin muistaa (ks. monisteen luku 5.3.4 valintakorjauksista) että jos aineistossa on paljon muuttujia, osa niiden välillä havaituista riippuvuuksista selittyy pelkällä sattumalla. Jos aineistosta etsitään "epätavallisia" riippuvuuksia ja sitten testataan niitä, saadut tilastollisesti merkitsevät tulokset johtuvat sattumasta huomattavasti todennäköisemmin kuin saatu p-arvo kertoisi.

5. (Vrt. monisteen teht. 5.2.) Kolikon harhattomuutta tutkitaan heittämällä sitä n kertaa ja kirjaamalla ylös kruunujen lukumäärä.

a) Formuloi huolellisesti asetelmaa kuvaava malli ja nollahypoteesi. Mikä on luonnollinen testisuure ja millaiset testisuureen arvot puhuvat nollahypoteesia vastaan?

b) Heittoja on $n = 10$ ja saadaan 7 kruunua. Laske vastaava p-arvo ja pohdi, voidaanko kolikkoa pitää harhattomana.

c) Muuttuvatko johtopäätöksesi, jos $n = 100$ ja saadaan 70 kruunua?

Ohje. Binomijakaumaan liittyviä todennäköisyyksiä voi laskea kätevästi netissä olevilla laskimilla, esim. http://www.jkauppi.fi/mathematics/binomial-calculator. c-kohdassa voi käyttää binomijakauman normaaliapproksimaatiota.

Ratkaisu 5

a) Merkitään K = kruunujen lukumäärä ja n = heittojen lukumäärä. K on binomijakautunut, eli $K \sim Bin(n, \theta)$ ja nollahypoteesi kolikon harhattomuudesta on $H_0 : \theta = \frac{1}{2}$.

Luonnollinen testisuure on K itse, ja koska $\mathrm{E}_{H_0}(K) = n\theta = \frac{n}{2}$, K:n suuret ja pienet arvot todistavat nollahypoteesia vastaan.

b) Lasketaan p-arvo:

$$\begin{aligned} p &= P_{H_0}\left\{|K - \frac{n}{2}| \geq |k - \frac{n}{2}|\right\} \\ &= P_{H_0}\left\{K \geq \frac{n}{2} + |k - \frac{n}{2}|\right\} + P_{H_0}\left\{K \leq \frac{n}{2} - |k - \frac{n}{2}|\right\} \end{aligned}$$

Koska binomijakauman odotusarvo on $n\theta$ ja varianssi $n\theta(1-\theta)$, normaaliapproksimaatiota jatkuvuuskorjauksella käyttäen saadaan:

$$\begin{aligned} &P_{H_0}\left\{K \geq \frac{n}{2} + |k - \frac{n}{2}|\right\} \\ &\approx 1 - P\left\{Z \leq \frac{\frac{n}{2} + |k - \frac{n}{2}| - \frac{1}{2} - \frac{n}{2}}{\sqrt{n \cdot \frac{1}{2} \cdot (1 - \frac{1}{2})}}\right\} \\ &= 1 - \Phi\left(\frac{|k - \frac{n}{2}| - \frac{1}{2}}{\frac{\sqrt{n}}{2}}\right) \end{aligned}$$

ja

$$\begin{aligned} &P_{H_0}\left\{K \leq \frac{n}{2} - |k - \frac{n}{2}|\right\} \\ &\approx 1 - P\left\{Z \leq -\frac{\frac{n}{2} - |k - \frac{n}{2}| + \frac{1}{2} - \frac{n}{2}}{\sqrt{n \cdot \frac{1}{2} \cdot (1\frac{1}{2})}}\right\} \\ &= \Phi\left(\frac{-|k - \frac{n}{2}| + \frac{1}{2}}{\frac{\sqrt{n}}{2}}\right) \\ &= 1 - \Phi\left(\frac{|k - \frac{n}{2}| - \frac{1}{2}}{\frac{\sqrt{n}}{2}}\right), \end{aligned}$$

joten kun $n = 10$ ja $k = 7$, p-arvoksi saadaan

$$\begin{aligned} p &= 2\left(1 - \Phi\left(\frac{|k - \frac{n}{2}| - \frac{1}{2}}{\frac{\sqrt{n}}{2}}\right)\right) \\ &= 2\left(1 - \Phi\left(\frac{|7 - \frac{10}{2}| - \frac{1}{2}}{\frac{\sqrt{10}}{2}}\right)\right) \\ &= 2\left(1 - \Phi\left(\frac{3}{\sqrt{10}}\right)\right) \approx 0.343 \end{aligned}$$

Oltaisiin voitu myös laskea tarkka arvo binomijakaumasta; tarkka arvo on noin 0.344, joten normaaliapproksimaatio tuottaa tässä riittävän tarkan arvion.

Vähintään näin poikkeavan tuloksen saamisen todennäköisyys silloin kun nollahypoteesi pätee, eli kolikko on harhaton, on yli 1/3, joten kolikon harhattomuutta ei tämän perusteella ole vielä syytä epäillä.

c) Kun $n = 100$ ja $k = 70$, p-arvoksi saadaan

$$\begin{aligned} p &= 2\left(1 - \Phi\left(\frac{|k - \frac{n}{2}| - \frac{1}{2}}{\frac{\sqrt{n}}{2}}\right)\right) \\ &= 2\left(1 - \Phi\left(\frac{|70 - \frac{100}{2}| - \frac{1}{2}}{\frac{\sqrt{100}}{2}}\right)\right) \\ &= 2(1 - \Phi(3.9)) \approx 9.62 \cdot 10^{-5} \end{aligned}$$

Vähintään näin poikkeavan tuloksen saamisen todennäköisyys silloin kun nollahypoteesi pätee, eli kolikko on harhaton, on alle 1/10000, joten on erittäin vahvoja perusteita epäillä onko kolikko todella harhaton.