MS-A0502 Todennäköisyyslaskennan ja tilastotieteen peruskurssi

6A Tilastollisen merkitsevyyden testaaminen

Jukka Kohonen

Matematiikan ja systeemianalyysin laitos
Perustieteiden korkeakoulu
Aalto-yliopisto

Lukuvuosi 2019–2020
Periodi II

Sisältö

Hypoteesin testaus pähkinänkuoressa

1. Jos data $\vec{X}$ on peräisin erään hypoteesin H_0 mukaisesta stokastisesta mallista, niin silloin siitä funktiolla t laskettu tunnusluku $t(\vec{X})$ noudattaa erästä jakaumaa J.
2. Lasketaan havaitusta datasta tunnusluku $t(\vec{x})$.
3. Onko arvo $t(\vec{x})$ "kovin yllättävä" tulos jakaumasta J? Jos on, niin hypoteesi H_0 ei vaikuta uskottavalta ja se "hylätään".

Huomautuksia.

1. Hypoteesin hylkäys ei tarjoa mitään vaihtoehtoista selitystä datalle. Se pitää tehdä erikseen.
2. Vaikka hypoteesi olisi oikeasti totta, se saatetaan hylätä, koska yllätyksiä sattuu (satunnaisuus!).
3. Ei ole selvää, millainen t:n arvo on "liian" yllättävä.

Mustekala Paul

Jalkapallon MM-kisoissa 2010 Paul ennusti voittajan oikein jokaiselle Saksan ottelulle.

Opponent	Tournament	Stage	Date	Prediction	Result	Outcome
Poland	Euro 2008	group stage	8 June 2008	Germany	2–0	Correct
Croatia	Euro 2008	group stage	12 June 2008	Germany[3][20]	1–2	Incorrect
Austria	Euro 2008	group stage	16 June 2008	Germany	1–0	Correct
Portugal	Euro 2008	quarter-finals	19 June 2008	Germany	3–2	Correct
Turkey	Euro 2008	semi-finals	25 June 2008	Germany	3–2	Correct
Spain	Euro 2008	final	29 June 2008	Germany[3]	0–1	Incorrect
Australia	World Cup 2010	group stage	13 June 2010	Germany[31]	4–0	Correct
Serbia	World Cup 2010	group stage	18 June 2010	Serbia[31]	0–1	Correct
Ghana	World Cup 2010	group stage	23 June 2010	Germany[31]	1–0	Correct
England	World Cup 2010	round of 16	27 June 2010	Germany[32]	4–1	Correct
Argentina	World Cup 2010	quarter-finals	3 July 2010	Germany[23]	4–0	Correct
Spain	World Cup 2010	semi-finals	7 July 2010	Spain[33]	0–1	Correct
Uruguay	World Cup 2010	3rd place play-off	10 July 2010	Germany	3–2	Correct

Onko poikkeuksellisen hyvä ennustustulos tilastollisesti merkitsevä, vai voidaanko se lukea tavanomaisen satunnaisvaihtelun piiriin?

https://en.wikipedia.org/wiki/Paul_the_Octopus

Nollahypoteesi H_0

Tilastollisen merkitsevyystestin lähtökohdaksi muotoillaan nollahypoteesi H_0, joka vastaa tilannetta, jossa mitään uutta tai yllättävää ei tarvita havaintojen selittämiseen.

Esim

H_0: Selvännäkijän ennustukset eivät ole arvauksia parempia

H_0: Uusi lääke ei ole lumelääkettä tehokkaampi

H_0: Salkunhoitajan rahaston tuotto ei ole pörssi-indeksiä parempi

Merkitsevyystestin vastahypoteesi H_1 on yleensä nollahypoteesin vastakohta.

Nollahypoteesi vs. data

Kuuluuko havaittu data tyypillisen satunnaisvaihtelun piiriin vai onko syytä epäillä nollahypoteesia?

Esimerkki (Kolikko)

Tasaiseksi väitettyä kolikkoa 50 kertaa heitettäessä saadaan 42 kruunaa.
H_0: Kruunan tn $\theta = 1/2$
H_1: Kruunan tn $\theta \neq 1/2$

Esimerkki (Kohinainen kanava)

Tiedonsiirtovirheiden väitetään olevan normaalijakautuneita parametreina $\mu = 0$ ja $\sigma = 3$. Kanavaa kerran testaamalla mitattiin virheeksi $x_1 = 4.8$.
H_0:$\mu = 0$
H_1:$\mu \neq 0$

Esimerkki (Laadunvalvonta)

Tukkukauppias väittää, että sen tomaateista enintään 5% on huonoja. Suuresta erästä poimittiin 50 tomaattia ja niistä 4 todettiin huonoiksi.
H_0:Huonolaatuisten osuus $\theta \leq 0.05$
H_1:Huonolaatuisten osuus $\theta > 0.05$

Testisuureen p-arvo

Havaitun datajoukon $\vec{x} = (x_1, \ldots, x_n)$ poikkeuksellisuutta analysoidaan laskemalla testisuure

$$t(\vec{x}) = t(x_1, \ldots, x_n),$$

joka tiivistää havaitut datapisteet yhdeksi reaaliluvuksi.

Testisuureen p-arvo on todennäköisyys, jolla nollahypoteesin mukaisen datalähteen ennakoidaan tuottavan ainakin niin poikkeavia testisuureen arvoja kuin $t(\vec{x})$.

p-arvo	**Tulkinta**
> 0.10	Havainto ei ole ristiriidassa H_0:n kanssa
≈ 0.05	Havainto todistaa jonkun verran H_0:aa vastaan
< 0.01	Havainto todistaa vahvasti H_0:aa vastaan

Esim. Kolikko

Tasaiseksi väitetty kolikko tuottaa 42 kruunaa 50 heitolla.
H_0: Kruunan tn $\theta = 1/2$
H_1: Kruunan tn $\theta \neq 1/2$
Testisuure = kruunien lukumäärä: $t(x) = 42$
$T = t(X) =$ "kruunien lkm stokastisessa mallissa"

$$f(x) \;=\; \mathbb{P}(T = x \,|\, H_0) \;=\; \binom{50}{x}\left(\frac{1}{2}\right)^x \left(1 - \frac{1}{2}\right)^{50-x}$$

Testisuureen odotusarvo $t_0 = \mathbb{E}(T \,|\, H_0) = 25$.

$$\begin{aligned} \text{p-arvo} \;&=\; \mathbb{P}(|T - t_0| \geq |t(x) - t_0| \,|\, H_0) \\ &=\; \mathbb{P}(|T - 25| \geq 17 \,|\, H_0) \\ &=\; \sum_{x=0}^{8} f(x) + \sum_{x=42}^{50} f(x) \;\approx\; 1.2 \times 10^{-6}. \end{aligned}$$

Havainto todistaa vahvasti H_0:aa vastaan.

Yllättävyyden määritelmästä

Miksi tarkastellaan “ainakin” yhtä poikkeavan arvon todennäköisyyttä — eikä esim. testisuureen arvon todennäköisyyttä sinänsä?

Esimerkki

Heitetään kolikkoa $n = 1000$ kertaa. Jos kolikko on reilu (H_0), niin testisuure $t(X)$ eli kruunien määrä noudattaa binomijakaumaa $\mathrm{Bin}(500, 1/2)$.

Havaitaan testisuureen arvo $t(x) = 502$. Binomijakauman mukaan $\Pr(t(X) = 502) \approx 0.0250$, aika pieni luku. Tulisiko meidän olla yllättyneitä ja epäillä kolikkoa epäreiluksi?

Ei ehkä kovin yllättyneitä, sillä myös $\Pr(t(x) = 500) \approx 0.0252$, aika pieni. Mikään yksittäinen arvo ei ole kovin todennäköinen.

Tavanomainen ratkaisu on olla “yllättynyt” vasta, kun olisi epätodennäköistä saada tämä $t(x)$ tai jotain vielä poikkeavampaa (testisuureen jakauman häntätodennäköisyys).

Esim. Kohinainen kanava

Tiedonsiirtovirheiden väitetään olevan normaalijakautuneita odotusarvona $\mu = 0$ ja keskihajontana 3. Havainto: $x_1 = 4.8$.
H_0: Odotusarvo $\mu = 0$
H_1: Odotusarvo $\mu \neq 0$
Testisuure = normitettu poikkeama nollahypoteesin mukaisesta odotusarvosta: $z(\vec{x}) = \frac{x_1 - 0}{3} = 1.6$

$$\text{p-arvo} \ = \ \mathbb{P}(|Z| \geq 1.6 \,|\, H_0) \ = \ 2\mathbb{P}(Z \geq 1.6 \,|\, H_0) \ \approx \ 11\%,$$

Havainto on selitettävissä tavanomaisella satunnaisvaihtelulla. Havainto ei puhu H_0:aa vastaan.

Sisältö

Esim. Laadunvalvonta

Tukkukauppiaan väitteen mukaan sen toimittamista tomaateista enintään 5% on huonolaatuisia. Suuresta erästä poimittiin satunnaisesti 50 tomaattia ja niistä 4 todettiin huonolaatuisiksi.
H_0: Huonojen osuus $\theta \leq 0.05$
H_1: Huonojen osuus $\theta > 0.05$
Testisuure: Huonojen lkm: $t(\vec{x}) = 4$
Datalähteen tuottamien testisuureen arvojen stokastinen malli:

$$\mathbb{P}_\theta(T = t) \;=\; f_\theta(t) \;=\; \binom{50}{t}\theta^t(1-\theta)^{50-t}$$

Poikkeavuus: testisuureen arvo poikkeaa *ylöspäin* odotusarvosta

$$\mathbb{P}_\theta\Big(T - \mathbb{E}_\theta(T) \geq t(\vec{x}) - \mathbb{E}_\theta(T) \Big) \;=\; \mathbb{P}_\theta(T \geq t(\vec{x})) \;=\; \sum_{t=4}^{50} f_\theta(t).$$

Ongelma: tn riippuu θ:sta. Valitaan suurin tn (miksi?).

$$\text{p-arvo} \;=\; \max_{\theta \leq 0.05} \mathbb{P}_\theta(T \geq t(\vec{x})) = \mathbb{P}_{0.05}(T \geq t(\vec{x})) = \sum_{t=4}^{50} f_{0.05}(t) \;\approx\; 24\%$$

Tilastollisen merkitsevyystestin vaiheet

Joskus asetetaan myös vastahypoteesi H_1, jolloin tarkoitus on tutkia, puoltaako havaittu datajoukko enemmän H_0:n vai H_1:n hyväksymistä.

1. Asetetaan nollahypoteesi H_0 ja mahdollisesti vastahypoteesi H_1.
2. Valitaan käytettävä testisuure ja lasketaan havaittua dataa $\vec{x}$ vastaava testisuureen arvo $t(\vec{x})$.
3. Määritetään datalähteen tuottamia testisuureen arvoja mallintavan satunnaismuuttujan T jakauma nollahypoteesin vallitessa.
4. Määritetään mitkä testisuureen arvot tulkitaan havaittua testisuureen arvoa poikkeuksellisemmiksi (puoltavat H_1:n hyväksymistä H_0:n sijaan) ja lasketaan testin p-arvo.
5. Tehdään johtopäätökset:
    - Pieni p-arvo $\implies$ H_0 hylätään
    - Suuri p-arvo $\implies$ H_0 jää voimaan

Testaaminen valitulla merkitsevyystasolla

Miten pieni p-arvo on riittävän pieni?

Tietyissä tilanteissa vaaditaan yksiselitteistä johtopäätöstä: testin pohjalta H_0 joko hyväksytään tai hylätään. Tällaisen testin pohjaksi valitaan merkitsevyystaso $\alpha \in (0, 1)$ ja johtopäätös muodostetaan seuraavasti:

- Jos p-arvo $\geq \alpha$, nollahypoteesi hyväksytään (jää voimaan),
- Jos p-arvo $< \alpha$, nollahypoteesi hylätään.

Usein on tapana valita merkitsevyystasoksi $\alpha = 1\%$ tai $\alpha = 5\%$

Sisältö

Nollahypoteesi ja p-arvo

Yhdistetty nollahypoteesi

Testausvirheet

Odotusarvon testaaminen suurille datajoukoille

Testausvirheet

Mikään ei takaa, että tehty johtopäätös olisi oikea.

		Tehty johtopäätös	
		H_0 hyväksytään	H_0 hylätään
Totuus	H_0 tosi	Oikea päätös	Hylkäysvirhe
	H_0 epätosi	Hyväksymisvirhe	Oikea päätös

Testaajan johtopäätös on aina arvaus. Hyvä arvaus on todennäköisesti oikein.

Tn-laskennan keinoin voidaan yrittää laskea, millä *todennäköisyydellä* sattuu hyväksymis- ja hylkäysvirheitä.

Tilastotieteessä käytetään myös nimiä "tyypin I virhe" (väärä hylkäys) ja "tyypin II virhe" (väärä hyväksyminen).

Testausvirheiden todennäköisyydet

$p(\vec{x})$ = testisuureen p-arvo datajoukolle $\vec{x}$
$\vec{X} = (X_1, \ldots, X_n)$ mallintaa datalähteen tuottamia arvoja ennen niiden havaitsemista $\implies$ $p(\vec{X})$ on satunnaisluku
Hylkäysvirheen todennäköisyys on

$$\mathbb{P}(H_0 \text{ hylätään} \mid H_0) \;=\; \mathbb{P}(p(\vec{X}) < \alpha \mid H_0)$$

Hyväksymisvirheen todennäköisyys on

$$\mathbb{P}(H_0 \text{ hyväksytään} \mid H_1) \;=\; \mathbb{P}(p(\vec{X}) \geq \alpha \mid H_1),$$

α	Hylkäysvirheen tn	Hyväksymisvirheen tn
Lähellä nollaa	Pieni	Suuri
Lähellä ykköstä	Suuri	Pieni

Fakta

Hylkäysvirheen tn $\leq \alpha$. *(Miksi?)*

Testausvirheiden tulkinta

Anni Aktiivi

- Käyttää merkitsevyystasoa $\alpha = 5\%$
- Hylkää useammin nollahypoteeseja
- On henkisesti varautunut siihen, että tietty osuus testien johtopäätöksistä on virheellisiä
- Tietää, että pitkällä tähtäyksellä hänen hylkäämistään nollahypoteeseista enintään 5% on virheellisesti hylätty (Hän ei kuitenkaan tiedä mitkä niistä.)
- Hyväksyy Villeä harvemmin virheellisesti nollahypoteeseja (ei tiedä miten usein)

Ville Varovainen

- Käyttää merkitsevyystasoa $\alpha = 1\%$
- Hylkää harvemmin nollahypoteeseja
- On henkisesti varautunut siihen, että tietty osuus testien johtopäätöksistä on virheellisiä
- Tietää, että pitkällä tähtäyksellä hänen hylkäämistään nollahypoteeseista enintään 1% on virheellisesti hylätty (Hän ei kuitenkaan tiedä mitkä niistä.)
- Hyväksyy Annia useammin virheellisesti nollahypoteeseja (ei tiedä miten usein)

Esim. Rikosoikeus

H_0: Epäilty on syytön
H_1: Epäilty on syyllinen
Havaittu data: Saatavilla oleva todistusaineisto
Hylkäysvirhe: Syytön tuomitaan
Hyväksymisvirhe: Syyllistä ei tuomita

Anni Aktiivi

- Käyttää merkitsevyystasoa $\alpha = 5\%$
- Langettaa useammin tuomioita
- Tuomitsee pitkällä tähtäyksellä $\leq 5\%$ syyttömiä
- Jättää Villeä harvemmin syyllisiä tuomitsematta

Ville Varovainen

- Käyttää merkitsevyystasoa $\alpha = 1\%$
- Langettaa harvemmin tuomioita
- Tuomitsee pitkällä tähtäyksellä $\leq 1\%$ syyttömiä
- Jättää Annia useammin syyllisiä tuomitsematta

Esim. Kolikko

Tasaiseksi väitettyä kolikkoa 10 kertaa heitettäessä havaitaan data $\vec{x} = (0, 0, 1, 0, 0, 0, 0, 0, 0, 0)$. Testaa väitettä 5% merkitsevyystasolla.

H_0: Kruunan tn $\theta = 0.5$,
H_1: Kruunan tn $\theta \neq 0.5$.

Testisuure: $t(\vec{x})$=kruunien lkm
Testisuureen stokastinen malli: $T = t(\vec{X})$

$$f_{H_0}(t) \;=\; \mathbb{P}(T = t \mid H_0) \;=\; \binom{10}{t} \left(\frac{1}{2}\right)^{10}$$

Havainnon $\vec{x}$ p-arvo

$$p(\vec{x}) \;=\; \mathbb{P}(\,|t(\vec{X}) - 5| \geq 4 \mid H_0) \;=\; \sum_{t=0}^{1} f_{H_0}(t) + \sum_{t=9}^{10} f_{H_0}(t) \;\approx\; 2.1\%.$$

Johtopäätös: Nollahypoteesi hylätään (5% merkitsevyystasolla).
Mitä osataan sanoa virhetodennäköisyyksistä?

Esim. Kolikko — Hylkäysvirheen tn

Testin p-arvot testisuureen funktiona:

# kruunat	0	1	2	3	4	5	6	7	8	9	10
p-arvo (%)	0.2	2.1	10.9	34.4	75.4	100	75.4	34.4	10.9	2.1	0.2

5% merkitsevyystasolla testin hylkäysalue on $\{0, 1, 9, 10\}$.
Hylkäysvirheen tn on

$$\mathbb{P}(t(\vec{X}) \in \{0, 1, 9, 10\} \mid H_0) \ = \ \sum_{t=0}^{1} f_{H_0}(t) + \sum_{t=9}^{10} f_{H_0}(t) \ \approx \ 2.1\%.$$

Esim. Kolikko — Hyväksymisvirheen tn

Testin p-arvot testisuureen funktiona:

# kruunat	0	1	2	3	4	5	6	7	8	9	10
p-arvo (%)	0.2	2.1	10.9	34.4	75.4	100	75.4	34.4	10.9	2.1	0.2

5% merkitsevyystasolla testin hylkäysalue on $\{0, 1, 9, 10\}$.
Hyväksymisvirheen tn on

$$\mathbb{P}(t(\vec{X}) \in \{2, 3, \ldots, 8\} \mid H_1) \ = \ ?$$

Ongelma: vastahypoteesi ($H_1 : \theta \neq 0.5$) ei määrää θ:n arvoa.
Ääritapaus $\theta \approx 0.5$ (esim. $\theta = 0.5001$), jolloin

$$\begin{aligned}\mathbb{P}(t(\vec{X}) \in \{2, 3, \ldots, 8\} \mid H_1) \ &\approx \ \mathbb{P}(t(\vec{X}) \in \{2, 3, \ldots, 8\} \mid H_0) \\ &= \ \sum_{t=2}^{8} f_{H_0}(t) \ \approx \ 97.9\%.\end{aligned}$$

Esim. Kahden tyyppisiä kolikoita

Tasaiseksi väitettyä kolikkoa 10 kertaa heitettäessä havaitaan data $\vec{x} = (0, 0, 1, 0, 0, 0, 0, 0, 0, 0)$. Tiedetään, että joko $\theta = 0.5$ tai $\theta = 0.9$. Testaa väitettä 5% merkitsevyystasolla.

> H_0: *Kruunan tn* $\theta = 0.5$,
> H_1: *Kruunan tn* $\theta = 0.9$.

Sama testisuure, sama p-arvo, sama johtopäätös (H_0 hylätään).

$$f_{H_1}(t) \;=\; \mathbb{P}(T = t \,|\, H_1) \;=\; \binom{10}{t} 0.9^t (1 - 0.9)^{10-t}$$

Hyväksymisvirheen tn:

$$\mathbb{P}(\, t(\vec{X}) \in \{2, 3, \ldots, 8\} \,|\, H_1) \;=\; \sum_{t=2}^{8} f_{H_1}(t) \;\approx\; 26\%.$$

Sisältö

Odotusarvon testisuure suurille datajoukoille

Datalähde tuottaa toisistaan riippumattomia ja samoin jakautuneita satunnaislukuja $X_1, X_2, \ldots, X_n$ odotusarvona (tuntematon) μ. Datalähteen väitetty odotusarvo on μ_0.

H_0: $\mu = \mu_0$
H_1: $\mu \neq \mu_0$

Jakauma tuntematon $\implies$ mahdoton testata?
Approksimatiivinen testi mahdollinen, jos paljon dataa.
Testisuure:

$$t(\vec{x}) \;=\; \frac{m(\vec{x}) - \mu_0}{\mathrm{sd}(\vec{x})/\sqrt{n}}.$$

Fakta

Suurille datajoukoille testisuureen stokastinen malli $t(\vec{X})$ noudattaa likimain normitettua normaalijakaumaa, jolloin

$$\text{p-arvo} \;\approx\; \mathbb{P}(|t(\vec{X})| \geq |t(\vec{x})| \mid H_0) \;\approx\; \mathbb{P}(|Z| \geq |t(\vec{x})|).$$

Esim. Kahviautomaatti

Kahviautomaatin on tarkoitus laskea jokaiseen kuppiin keskimäärin 10.0 cl kahvia. Kahviautomaatin toimintaa testattiin valuttamalla automaatista 30 kupillista ja mittamalla kahvin määrät kupeissa.

Mittauksessa havaittiin arvot (cl):
11.05 9.65 10.93 9.46 10.27 10.02 10.07 10.74 11.15 10.40 10.12
11.20 10.07 10.27 9.99 9.80 10.83 10.21 11.26 10.11 10.49 10.10
10.15 11.02 10.00 11.68 10.51 11.20 11.29 10.15

Onko kahviautomaatti oikein kalibroitu?

Mittausdatan $\vec{x}$ keskiarvo on $m(\vec{x}) = 10.473$, joka poikkeaa tavoitearvosta $\mu_0 = 10.0$.

Onko poikkeama tilastollisesti merkitsevä?

Esim. Kahviautomaatti

11.05 9.65 10.93 9.46 10.27 10.02 10.07 10.74 11.15 10.40 10.12 11.20
10.07 10.27 9.99 9.80 10.83 10.21 11.26 10.11 10.49 10.10 10.15 11.02
10.00 11.68 10.51 11.20 11.29 10.15

Datajoukon keskiarvo $m(\vec{x}) = 10.473$, keskihajonta $\mathrm{sd}(\vec{x}) = 0.563$

H_0: $\mu = 10.0$
H_1: $\mu \neq 10.0$

Havaitun datajoukon testisuure:

$$t(\vec{x}) \;=\; \frac{m(\vec{x}) - \mu_0}{\mathrm{sd}(\vec{x})/\sqrt{n}} \;=\; \frac{10.473 - 10.0}{0.563/\sqrt{30}} \;=\; 4.60$$

Koska datajoukon koko $n = 30$ on melko suuri, käytetään suuren datajoukon testiä:

$$\text{p-arvo} \;\approx\; \mathbb{P}(|t(\vec{X})| \geq |t(\vec{x})| \mid H_0) \;\approx\; \mathbb{P}(|Z| \geq 4.60) \;\approx\; 4.2 \times 10^{-6}$$

Johtopäätös: Hyvin pieni p-arvo puoltaa vahvasti H_0:n hylkäämistä.

Hypoteesin testaus laajemmin

Edellä testattiin *odotusarvon* hypoteeseja (esim. $p = 1/2$ tai $\mu = 10.0$) testausta *vahvoilla yksinkertaistavilla oletuksilla* (esim. "paljon dataa, testisuure normaali").

Yleisemmin tilastotieteessä tutkitaan muitakin kysymyksiä, esim.

- Muiden hypoteesien testaus. Esim. odotusarvo tunnetaan, mutta kysytään onko datalähteen keskihajonta $\sigma = \sigma_0$ vai jotain muuta? $\longrightarrow \chi^2$-testi
- Heikommat oletukset. Esim. testataan odotusarvoa, dataa on vähän emmekä tunne lähteen keskihajontaa. Testisuure ei olekaan normaalijakautunut. $\longrightarrow$ Studentin t-testi

Näissäkin sovelletaan samaa perusasetelmaa.

1. Muotoillaan hypoteesi H (oletus datalähteen jakaumasta).
2. Muotoillaan testisuure, päätellään sen jakauma (jos H totta).
3. Tutkitaan, sopiiko testisuureen havaittu arvo ko. jakaumaan.

Näistä lisää esim. kurssilla Statistical inference MS-C1620.

Torstaina 5.12.2019 klo 10–12 kertausluento. (Perjantaina ei luentoa, koska itsenäisyyspäivä.)

Kertausluennolle saa ja kannattaa tuoda mukanaan kysymyksiä kurssilla käsitellyistä aiheista. Kysymyksiä voi lähettää myös etukäteen sähköpostilla.