Ilkka Mellin

Tilastolliset menetelmät

Osa 3: Tilastolliset testit

Testit laatueroasteikollisille muuttujille

Testit laatueroasteikollisille muuttujille

>> Laatueroasteikollisten muuttujien testit

Testi suhteelliselle osuudelle

Suhteellisten osuuksien vertailutesti

Testit laatueroasteikollisille muuttujille 1/2

- Tarkastelemme seuraavia testejä *laatueroasteikollisille* muuttujille:
  - **Testi suhteelliselle osuudelle**
  - **Suhteellisten osuuksien vertailutesti**
- On syytä huomata, että testejä saa – ja on usein myös järkevää – käyttää *järjestys*-, *välimatka*- ja *suhde-asteikollisille muuttujille*.
- Mitta-asteikot: ks. lukua **Tilastollisten aineistojen kerääminen ja mittaaminen.**

Testit laatueroasteikollisille muuttujille 2/2

- Testit ovat *parametrisia testejä*, joissa testauksen kohteena on **Bernoulli-jakauman odotusarvoparametri**.
- *Testi suhteelliselle osuudelle* on **yhden otoksen testi**.
- *Suhteellisten osuuksien vertailutesti* on **kahden otoksen testi**.

Testit laatueroasteikollisille muuttujille

Laatueroasteikollisten muuttujien testit

>> Testi suhteelliselle osuudelle

Suhteellisten osuuksien vertailutesti

Testausasetelma 1/3

- Olkoon A perusjoukon S *tapahtuma* ja olkoot

$$\Pr(A) = p$$

$$\Pr(A^c) = 1 - p = q$$

- Määritellään satunnaismuuttuja X:

$$X = \begin{cases} 1, & \text{jos } A \text{ sattuu} \\ 0, & \text{jos } A \text{ ei satu} \end{cases}$$

- Tällöin $X \sim \text{Bernoulli}(p)$ ja

$$\Pr(X = 1) = p$$

$$\Pr(X = 0) = 1 - p = q$$

Testausasetelma 2/3

- Oletetaan, että tapahtuma A on muotoa

  A = ”Perusjoukon alkiolla on ominaisuus P”

- Tällöin

  $p = \Pr(A)$

  on todennäköisyys poimia perusjoukosta S satunnaisesti alkio, jolla on ominaisuus P.

- Jos perusjoukko S on *äärellinen*, niin todennäköisyys p kuvaa niiden perusjoukon S alkioiden *suhteellista osuutta*, joilla on ominaisuus P.

Testausasetelma 3/3

- Olkoon $X_1 , X_2 , \ldots , X_n$ *yksinkertainen satunnaisotos* perusjoukosta *S*, joka noudattaa *Bernoulli-jakaumaa*

  Bernoulli(p)

- Asetetaan Bernoulli-jakauman *parametrille p nollahypoteesi*

  $\mathrm{H}_0 : p = p_0$

- Testausongelma:

  Ovatko havainnot *sopusoinnussa* nollahypoteesin H_0 kanssa?

- Ongelman ratkaisuna on **testi suhteelliselle osuudelle**.

Hypoteesit

- *Yleinen hypoteesi* H :

  (1) Havainnot $X_i \sim \text{Bernoulli}(p)$, $i = 1, 2, \text{K}, n$, jossa

  $p = \Pr(A)$, $A \subset S$

  (2) Havainnot $X_1, X_2, \ldots, X_n$ ovat *riippumattomia*

- *Nollahypoteesi* H_0 :

  $\text{H}_0 : p = p_0$

- *Vaihtoehtoinen hypoteesi* H_1 :

  $\left.\begin{array}{l} \text{H}_1 : p > p_0 \\ \text{H}_1 : p < p_0 \end{array}\right\}$ 1-suuntaiset vaihtoehtoiset hypoteesit

  $\text{H}_1 : p \neq p_0$ 2-suuntainen vaihtoehtoinen hypoteesi

Parametrien estimointi

- Olkoon *f tapahtuman A frekvenssi* siinä *n*-kertaisessa *toistokokeessa*, jota riippumattomien havaintojen poimiminen Bernoulli-jakaumasta merkitsee.
- Tällöin tapahtuman *A suhteellinen frekvenssi* eli *osuus*

$$\hat{p} = f / n$$

on *harhaton estimaattori* Bernoulli-jakauman parametrille

$$\mathrm{E}(X_i) = p \text{ , } i = 1, 2, \ldots, n$$

- Huomaa, että frekvenssi *f* noudattaa *binomijakaumaa* parametrein *n* ja *p*:

$$f = \sum_{i=1}^{n} X_i \sim \mathrm{Bin}(n, p)$$

Testisuure ja sen jakauma

- Määritellään **testisuure**

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

- *Jos nollahypoteesi*

$$\mathrm{H}_0 : p = p_0$$

pätee, niin testisuure z noudattaa suurissa otoksissa *approksimatiivisesti standardoitua normaalijakaumaa*:

$$z \;\; {}_{a} \;\; \mathrm{N}(0,1)$$

- Approksimaatio on tavallisesti *riittävän hyvä*, jos

$$n\hat{p} \geq 10 \text{ ja } n(1-\hat{p}) \geq 10$$

Testi suhteelliselle osuudelle

Testisuureen jakauma nollahypoteesin H_0 pätiessä: Perustelu

- Oletetaan, että testin yleinen hypoteesi H ja nollahypoteesi H_0 pätevät:

$$X_1, X_2, \ldots, X_n \perp$$

$$X_i \quad \text{Bernoulli}(p_0), i = 1, 2, \mathrm{K}, n$$

- Tällöin (ks. monisteen **Todennäköisyyslaskenta** lukua **Stokastiikan konvergenssikäsitteet ja raja-arvolauseet** sekä lukuja **Otokset ja otosjakaumat** ja **Väliestimointi**):

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{f}{n} \quad {}_a \quad \mathrm{N}\left(p_0, \frac{p_0(1-p_0)}{n} \right)$$

jolloin

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \quad {}_a \quad \mathrm{N}(0,1)$$

Testi suhteelliselle osuudelle: Testisuure *z* mittaa tilastollista etäisyyttä

- Testisuure

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

mittaa parametrin p estimaatin $\hat{p}$ ja nollahypoteesin $H_0 : p = p_0$ kiinnittämän parametrin p arvon p_0 *tilastollista etäisyyttä*.

- *Mittayksikkönä* on erotuksen $\hat{p} - p_0$ standardipoikkeaman

$$\sqrt{\frac{p(1-p)}{n}}$$

estimaattori, joka on määrätty olettaen, että nollahypoteesi H_0 pätee.

Testi

- Testisuureen

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

normaaliarvo = 0, koska *nollahypoteesin* H_0 *pätiessä*

$$E(z) = 0$$

- Siten itseisarvoltaan *suuret* testisuureen *z* arvot viittaavat siihen, että *nollahypoteesi* H_0 *ei päde*.
- *Nollahypoteesi* H_0 *hylätään*, jos testin *p*-arvo on *kyllin pieni*.
- Hylkäysalueen valinta ja *p*-arvon määrääminen: ks. lukua **Tilastollinen testaus.**

Testit laatueroasteikollisille muuttujille

Laatueroasteikollisten muuttujien testit

Testi suhteelliselle osuudelle

>> Suhteellisten osuuksien vertailutesti

Testausasetelma 1/4

- Olkoon A perusjoukon S_k , $k = 1, 2$ *tapahtuma* ja olkoot

  $\Pr(A) = p_k$

  $\Pr(A^c) = 1 - p_k = q_k$

- Määritellään satunnaismuuttujat X_k , $k = 1, 2$:

$$X_k = \begin{cases} 1, \text{ jos } A \text{ tapahtuu perusjoukossa } S_k \\ 0, \text{ jos } A \text{ ei tapahdu perusjoukossa } S_k \end{cases}$$

- Tällöin $X_k \sim \text{Bernoulli}(p_k)$, $k = 1, 2$ ja

  $\Pr(X_k = 1) = p_k$

  $\Pr(X_k = 0) = 1 - p_k = q_k$

Testausasetelma 2/4

- Oletetaan, että tapahtuma A on muotoa

  A = ”Perusjoukon alkiolla on ominaisuus P”

- Tällöin

  $p_k = \Pr(A)$

  on todennäköisyys poimia perusjoukosta S_k , $k = 1, 2$ satunnaisesti alkio, jolla on ominaisuus P.

- Jos perusjoukko S_k , $k = 1, 2$ on *äärellinen*, niin todennäköisyys p_k kuvaa niiden perusjoukon S_k alkioiden *suhteellista osuutta*, joilla on ominaisuus P.

Testausasetelma 3/4

- Olkoon

  $$X_{11}, X_{21}, \mathrm{K}, X_{n_1 1}$$

  yksinkertainen satunnaisotos perusjoukosta S_1, joka noudattaa *Bernoulli-jakaumaa*

  $$\text{Bernoulli}(p_1)$$

- Olkoon

  $$X_{12}, X_{22}, \mathrm{K}, X_{n_2 2}$$

  yksinkertainen satunnaisotos perusjoukosta S_2, joka noudattaa *Bernoulli-jakaumaa*

  $$\text{Bernoulli}(p_2)$$

- Olkoot otokset lisäksi toisistaan *riippumattomia*.

Testausasetelma 4/4

- Asetetaan Bernoulli-jakaumien *parametreille* p_1 ja p_2 *nollahypoteesi*

$$H_0 : p_1 = p_2 = p$$

- Testausongelma:

  Ovatko havainnot *sopusoinnussa* hypoteesin H_0 kanssa?

- Ongelman ratkaisuna on **suhteellisten osuuksien vertailutesti**.

Yleinen hypoteesi

- *Yleinen hypoteesi* H :

  (1) Havainnot X_{i1} Bernoulli(p_1), $i = 1, 2, \mathrm{K}, n_1$, jossa

  $p_1 = \Pr(A), A \subset S_1$

  (2) Havainnot X_{j2} Bernoulli(p_2), $j = 1, 2, \mathrm{K}, n_2$, jossa

  $p_2 = \Pr(A), A \subset S_2$

  (3) Havainnot X_{i1} ja X_{j2} ovat *riippumattomia* kaikille *i* ja *j*

- Huomautus:

  Oletus (3) sisältää *kolme riippumattomuusoletusta*:

  - Havainnot ovat riippumattomia otoksien 1 ja 2 *sisällä*.
  - Havainnot ovat riippumattomia otoksien 1 ja 2 *välillä*.

Nollahypoteesi ja vaihtoehtoiset hypoteesit

- *Nollahypoteesi* H_0 :

  $H_0 : p_1 = p_2 = p$

- *Vaihtoehtoinen hypoteesi* H_1 :

  $H_1 : p_1 > p_2$
  $H_1 : p_1 < p_2$
  } 1-suuntaiset vaihtoehtoiset hypoteesit

  $H_1 : p_1 \neq p_2$ 2-suuntainen vaihtoehtoinen hypoteesi

Parametrien estimointi

- Olkoon f_k *tapahtuman A frekvenssi* siinä n_k-kertaisessa toistokokeessa, jota riippumattomien havaintojen poimiminen Bernoulli-jakaumasta k merkitsee, $k = 1, 2$.
- Tällöin tapahtuman *A suhteellinen frekvenssi* eli *osuus*

$$\hat{p}_k = f_k / n_k \ , k = 1, 2$$

on *harhaton estimaattori* Bernoulli-jakauman parametrille

$$p_k = \mathrm{E}(X_{ik}) \ , i = 1, 2, \ldots , n_k \ , k = 1, 2$$

- Huomaa, että frekvenssi f_k noudattaa *binomijakaumaa* parametrein n_k ja p_k:

$$f_k = \sum_{i=1}^{n_k} X_{ik} \sim \mathrm{Bin}(n_k , p_k) \ , k = 1, 2$$

Yhdistetty otos

- *Jos nollahypoteesi* $\mathrm{H}_0 : p_1 = p_2 = p$ *pätee*, voidaan otokset *yhdistää* ja parametrin *p harhaton estimaattori* on tapahtuman *A* suhteellinen frekvenssi yhdistetyssä otoksessa:

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{f_1 + f_2}{n_1 + n_2}$$

- *Jos nollahypoteesi* H_0 *pätee*, niin

$$\mathrm{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}$$

$$= p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Testisuure ja sen jakauma

- Määritellään **testisuure**

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- *Jos nollahypoteesi*

$$\mathrm{H}_0 : p_1 = p_2 = p$$

pätee, niin testisuure *z* noudattaa suurissa otoksissa *approksimatiivisesti standardoitua normaalijakaumaa*:

$$z \;\; \underset{a}{} \;\; \mathrm{N}(0,1)$$

- Approksimaatio on tavallisesti *riittävän hyvä*, jos

$$n_1\hat{p}_1 \geq 5\,,\; n_1(1-\hat{p}_1) \geq 5\,,\; n_2\hat{p}_2 \geq 5\,,\; n_2(1-\hat{p}_2) \geq 5$$

Testisuureen jakauma nollahypoteesin H_0 pätiessä: Perustelu 1/3

- Oletetaan, että testin yleinen hypoteesi H ja nollahypoteesi H_0 pätevät:

$$X_{11}, X_{21}, \K, X_{n_1 1}, X_{12}, X_{22}, \K, X_{n_2 2} \perp$$

$$X_{i1} \quad \text{Bernoulli}(p), \; i = 1, 2, \K, n_1$$

$$X_{j2} \quad \text{Bernoulli}(p), \; j = 1, 2, \K, n_2$$

- Tällöin (ks. monisteen **Todennäköisyyslaskenta** lukua **Stokastiikan konvergenssikäsitteet ja raja-arvolauseet** sekä lukuja **Otokset ja otosjakaumat** ja **Väliestimointi**):

$$\hat{p}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{i1} = \frac{f_1}{n_1} \quad \overset{a}{} \; \text{N}\left(p, \frac{p(1-p)}{n_1} \right)$$

$$\hat{p}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} X_{j2} = \frac{f_2}{n_2} \quad \overset{a}{} \; \text{N}\left(p, \frac{p(1-p)}{n_2} \right)$$

Testisuureen jakauma nollahypoteesin H_0 pätiessä: Perustelu 2/3

- Koska $\hat{p}_1 \perp \hat{p}_2$, niin

$$Y = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad {}_a \ \mathrm{N}(0,1)$$

- Koska todennäköisyys p on *tuntematon*, satunnaismuuttujan Y lauseke on testisuureena *epäoperationaalinen*.

Testisuureen jakauma nollahypoteesin H_0 pätiessä: Perustelu 3/3

- Jos satunnaismuuttujan *Y* lausekkeessa todennäköisyys *p* korvataan *otossuureella*

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{f_1 + f_2}{n_1 + n_2}$$

saadaan *testisuure*

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

joka *nollahypoteesin* H_0 *pätiessä noudattaa suurissa otoksissa standardoitua normaalijakaumaa* N(0, 1):

$$z \sim_a \mathrm{N}(0, 1)$$

- Todistus sivuutetaan.

Testisuure *z* mittaa tilastollista etäisyyttä

- Testisuure

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

  mittaa mittaa tapahtuman *A* otoksista 1 ja 2 määrättyjen suhteellisten frekvenssien *tilastollista etäisyyttä*.

- *Mittayksikkönä* on erotuksen $\hat{p}_1 - \hat{p}_2$ standardipoikkeaman

$$\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

  estimaattori, joka on määrätty olettaen, että nollahypoteesi H_0 pätee.

Testi 1/2

- Testisuureen

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

normaaliarvo = 0, koska *nollahypoteesin* H_0 *pätiessä*

$$E(z) = 0$$

- Siten itseisarvoltaan *suuret* testisuureen *z* arvot viittaavat siihen, että *nollahypoteesi* H_0 *ei päde*.
- *Nollahypoteesi* H_0 *hylätään*, jos testin *p*-arvo on *kyllin pieni*.

Testi 2/2

- Hylkäysalueen valinta ja *p*-arvon määrääminen:
  ks. lukua **Tilastollinen testaus.**