

# Sovellettu todennäköisyyslaskenta B

Antti Rasila

22. marraskuuta 2007

## 1 Epäparametrisia testejä (jatkoa)

- $\chi^2$ -riippumattomuustesti

## 2 Johdatus regressioanalyysiin

- Regressiomalli
- Regressio-ongelma
- Pienimmän neliösumman menetelmä

- $\chi^2$ -riippumattomuustestillä testataan, ovatko perusjoukon alkion tekijät  $A$  ja  $B$  riippumattomia.
- Käytännössä testin suorittaminen muistuttaa  $\chi^2$ -homogeenisuustestiä.
- Yleinen hypoteesi: Oletetaan, että perusjoukosta on poimittu yksinkertainen satunnaisotos ja havaintoyksiköt voidaan luokitella ristiin tekijöiden  $A$  ja  $B$  suhteen.

$H_0$ : Tekijät  $A$  ja  $B$  ovat riippumattomia.

$H_1$ : Tekijät  $A$  ja  $B$  eivät ole riippumattomia.

- Testi suoritetaan seuraavasti: valitaan havainnoille toisensa poissulkevat *luokitukset* tekijöiden  $A$  ja  $B$  suhteen.
- Luokitellaan havainnot tekijöiden  $A$  ja  $B$  suhteen, ja määrätään *havaitut luokkafrekvenssit*.

## $\chi^2$ -riippumattomuustestin suorittaminen 2/4

Kuten  $\chi^2$ -homogeenisuustestin tapauksessa, nämä voidaan esittää taulukkona:

	1	2	...	$c$	$\Sigma$
1	$O_{11}$	$O_{12}$	...	$O_{1c}$	$R_1$
2	$O_{21}$	$O_{22}$	...	$O_{2c}$	$R_2$
...	...	...	...	...	...
$r$	$O_{r1}$	$O_{r2}$	...	$O_{rc}$	$R_r$
$\Sigma$	$C_1$	$C_2$	...	$C_c$	$n$

Tässä  $r$  on A-luokkien lukumäärä,  $c$  B-luokkien lukumäärä,  $O_{ij}$  havaittu frekvenssi luokassa, jonka määrää A-luokka  $i$  ja B-luokka  $j$ ,  $R_i$  havaittu frekvenssi A-luokassa  $i$ ,  $C_j$  havaittu frekvenssi B-luokassa  $j$  ja  $n$  havaintojen kokonaislukumäärä.

## $\chi^2$ -riippumattomuustestin suorittaminen 3/4

- Erityisesti pätee, että

- (i)  $\sum_{j=1}^c O_{ij} = n_i$ ,
- (ii)  $\sum_{i=1}^r O_{ij} = C_j$ , ja
- (iii)

$$\sum_{i=1}^r \sum_{j=1}^c O_{ij} = \sum_{i=1}^r R_i = \sum_{j=1}^c C_j = n.$$

- Määrätään tehdyn riippumattomuusoletuksen mukaiset *odotetut luokkafrekvenssit*  $E_{ij}$  yhtälöillä:

$$E_{ij} = nP_{ij} = \frac{R_i C_j}{n}, \quad i = 1, \dots, r, \quad j = 1, \dots, c.$$

- Kuten edellä, näistä voidaan muodostaa taulukko.

## $\chi^2$ -riippumattomuustestin suorittaminen 4/4

- Verrataan havaittuja ja odotettuja luokkafrekvenssejä toisiinsa  $\chi^2$ -testisuurella:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

- Jos nollahypoteesi pätee, testisuure noudattaa suurissa otoksissa approksimatiivisesti  $\chi^2$ -jakaumaa vapausastein  $f = (r - 1)(c - 1)$ , missä  $r$  on A-luokkien ja  $c$  B-luokkien lukumäärä.
- Approksimaatio on tavallisesti riittävän hyvä, jos  $E_{ij} > 1$  ja keskimääräiset frekvenssit  $R_j/c > 5$  ja  $C_j/r > 5$ .
- Testisuureen normaaliarvo eli odotusarvo nollahypoteesin pätiessä on  $E(\chi^2) = f$ .
- Normaaliarvoa merkitsevästi suuremmat  $\chi^2$ -testisuureen arvot viittaavat siihen, että nollahypoteesi ei päde.

## Esimerkki 1/2

- Tehtaassa pyrittiin parantamaan tuotteen laatua. Ennen muutostöitä todettiin tarkastuksessa, että 80 tuotteen joukosta 65 oli täysin virheettömiä ja muutosten jälkeen 110 tuotteen joukosta 98 oli täysin virheettömiä. Paraniko tuotteiden laatu?
- Valitaan hypoteesiksi  $H_0$ : Muutostöillä ei ollut vaikutusta tuotteen laatuun (eli virheellisyys/virheettömyys on riippumatonta muutostöistä).
- Vaihtoehtoinen hypoteesi  $H_1$ : Muutostöillä oli vaikutusta.
- Luokitellaan tekijöiden (A) Muutostyöt ennen/jälkeen ( $=1/2$ ) ja (B) virheetön/virheellinen ( $=1/2$ ).
- Muodostetaan nelikenttä

	$A_1$	$A_2$	$\Sigma$
$B_1$	65	98	163
$B_2$	15	12	27
$\Sigma$	80	110	190



## Esimerkki 2/2

- Odotetut frekvenssit  $E_{ij}$ :

	$A_1$	$A_2$	$\Sigma$
$B_1$	68.63	94.37	163
$B_2$	11.37	15.63	27
$\Sigma$	80	110	190

- $\chi^2$ -testisuure:

$$\chi^2 = \frac{(65 - 68.64)^2}{68.64} + \frac{(98 - 94.37)^2}{94.37} + \frac{(15 - 11.37)^2}{11.37} + \frac{(12 - 15.63)^2}{15.63} \approx 2.335$$

- Testataan merkitsevyystasolla  $\alpha = 0.05$ . Käytetään  $\chi^2$ -jakaumaa vapausasteilla  $f = (2 - 1)(2 - 1) = 1$ .
- Hylkäysalueeksi saadaan  $(3.84, \infty)$ , joten nollahypoteesi hyväksytään merkitsevyystasoilla 0.05. Muutostöillä ei ollut tilastollisesti merkitsevää vaikutusta tuotteen laatuun.

- Oletetaan, että haluamme selittää jonkin *selitettävän muuttujan* havaittujen arvojen vaihtelun *selittävien muuttujien* havaittujen arvojen vaihtelun avulla.
- *Regressioanalyysissä* selitettävän muuttujan tilastolliselle riippuvuudelle selittävistä muuttujista pyritään rakentamaan tilastollinen malli, jota kutsutaan *regressiomalliksi*.
- Regressioanalyysin mahdollisia tavoitteita ovat:
  - (i) Selitettävän muuttujan ja selittävien muuttujien *tilastollisen riippuvuuden luonteen kuvaaminen*.
  - (ii) Selitettävän muuttujan arvojen ennustaminen.

- Regressiomallissa

$$y_j = f(x_j; \beta) + \varepsilon_j, j = 1, 2, \dots, n$$

on seuraavat osat:

$y_j$  = *Selitettävän muuttujan* havaittu arvo havaintoyksikössä  $j$ .

$x_j$  = *Selittävän muuttujan* havaittu arvo havaintoyksikössä  $j$ .

$\beta$  = Tuntematon ei-satunnainen *parametri*.

$\varepsilon_j$  = Satunnainen *virhetermi* havaintoyksikössä  $j$ .

- Regressioanalyysissä pyritään valitsemaan regressiomallin parametrin  $\beta$  arvo siten, että kaikista virhetermeistä  $\varepsilon_j$  tulee samanaikaisesti mahdollisimman pieniä.
- Pyritään siis valitsemaan parametri  $\beta$  siten, että käyrä

$$y = f(x; \beta)$$

kulkisi mahdollisimman läheltä jokaista havaintopistettä

$$(x_j, y_j) \in \mathbb{R}^2, j = 1, 2, \dots, n.$$

- Erään ratkaisun tähän käyränsovitusongelmaan tarjoaa *pienimmän neliösumman menetelmä*.

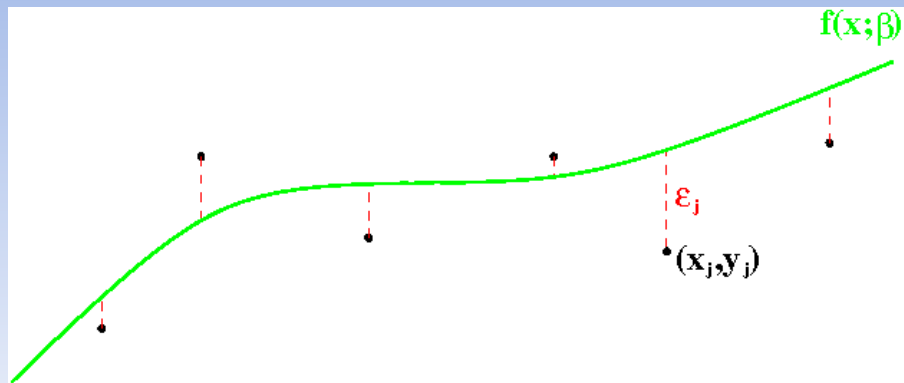
- Pienimmän neliösumman menetelmässä pyritään minimoimaan regressiomallin

$$y_j = f(x_j; \beta) + \varepsilon_j, j = 1, 2, \dots, n$$

virhetermien  $\varepsilon_j$  neliöden summaa muuttamalla parametrin  $\beta$  arvoa eli funktiota

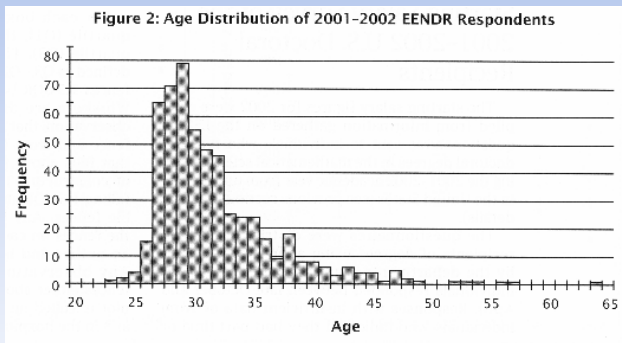
$$F(\beta) = \sum_{j=1}^n \varepsilon_j^2 = \sum_{j=1}^n (y_j - f(x_j; \beta))^2.$$

- Optimaalinen  $\beta$ :n arvo on parametrin  $\beta$  *pienimmän neliösumman estimaatti* eli *PNS-estimaatti*.



PNS-suoran sovittaminen: Kuvassa vihreällä parametr(e)ista  $\beta$  riippuva sovitettava funktio  $f(x; \beta)$  jollakin parametrin arvolla, datapisteet  $(x_j, y_j)$  ja vastaavat virhetermit  $\epsilon_j$ .

Eräessä tutkimuksessa selvitettiin tohtorintutkinon matematiikassa suorittavien ikää väitösvuotena. Tulokset olivat seuraavat:



## Esimerkki 2/3

- Sovitetaan dataan muotoa

$$f(x; \beta) = \beta_1 x^2 e^{-\beta_2 x}$$

oleva funktio.

- Tässä tapauksessa sovituksen tekeminen käsin ei ole helppoa. Tietokoneella (MATLAB) se kuitenkin onnistuu.

- Muodostetaan aluksi mallifunktio:

```
f=inline('beta(1)*x.^2.* exp(-beta(2)*x)', 'x', 'beta');
```

- Minimoitava funktio  $F: \mathbb{R}^2 \rightarrow \mathbb{R}_+$  on

$$F(\beta) = \sum_{j=1}^n (y_j - f(x_j; \beta))^2,$$

eli MATLAB:illa

```
fobj=inline('norm(feval(f,x,beta)-y)', 'beta');
```

missä  $x$  ja  $y$  ovat datan sisältävät vektorit.



## Esimerkki 3/3

- Minimointi voidaan suorittaa käskyllä `fminsearch`, eli `beta=fminsearch(fobj,beta0)`, missä `beta0` on minimin hakemisessa käytettävä alkuarvaus.
- Tulokseksi saadaan sovitus:

