

02402: Introduktion til Statistik

Forelæsning 10: Inferens for andele (forholdstal)

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Overview

- 1 Introduktion
- 2 Konfidensinterval for én andel
 - Stikprøvestørrelse, forsøgsplanlægning
- 3 Hypotesetest for én andel
- 4 Konfindensinterval og hypotesetest for to andele
- 5 Hypotesetest for flere andele
- 6 Statistik for antalstabeller

Overblik

- 1 Introduktion
- 2 Konfidensinterval for én andel
 - Stikprøvestørrelse, forsøgsplanlægning
- 3 Hypotesetest for én andel
- 4 Konfindensinterval og hypotesetest for to andele
- 5 Hypotesetest for flere andele
- 6 Statistik for antalstabeller

Forskellige analyse/data-situationer i 02402

Middelværdi for kvantitative data

- Hypotesetest/CI for én middelværdi (én stikprøve) (kapitel 3)
- Hypotesetest/CI for to middelværdier (to stikprøver) (kapitel 3)
- Hypotesetest/CI for flere middelværdier (K stikprøver) (kapitel 8, næste uge)

I dag: Andele (forholdstal, engelsk: *proportions*)

- Hypotetest/CI for én andel
- Hypotetest/CI for to andele
- Hypotesetest for flere andele
- Hypotesetest test for “multi-kategoriske” andele/tabeller

Estimation af andele

- Estimation af en andel (sandsynlighed) fås ved at observere antallet af gange (x) en hændelse har indtruffet ud af n (uafhængige) forsøg:

$$\hat{p} = \frac{x}{n}$$

Bemærk:

- $\hat{p} \in [0; 1]$.
- \hat{p} er en stokastisk variabel. Gentagelser af forsøget kan give forskellige udfald.

Overview

- 1 Introduktion
- 2 Konfidensinterval for én andel
 - Stikprøvestørrelse, forsøgsplanlægning
- 3 Hypotesetest for én andel
- 4 Konfidensinterval og hypotesetest for to andele
- 5 Hypotesetest for flere andele
- 6 Statistik for antalstabeller

Konfidensinterval for én andel

Method 7.3

Hvis stikprøven er **stor**, så er $(1 - \alpha)100\%$ konfidensintervallet for p givet ved:

$$\frac{x}{n} - z_{1-\alpha/2} \sqrt{\frac{\frac{x}{n}(1-\frac{x}{n})}{n}} < p < \frac{x}{n} + z_{1-\alpha/2} \sqrt{\frac{\frac{x}{n}(1-\frac{x}{n})}{n}}$$

Hvordan?

Følger af at approksimere binomialfordelingen med normalfordelingen.

Tommelfingerregel

Antag $X \sim \text{binom}(n, p)$. Normalfordelingen er en god tilnærmelse til binomialfordelingen hvis np og $n(1-p)$ (forventet antall succeser og fiaskoer) begge er større 15.

Konfidensinterval for én andel

Middelværdi og varians i binomialfordelingen, kapitel 2.21

$$\begin{aligned} E(X) &= np \\ \text{Var}(X) &= np(1-p) \end{aligned}$$

Altså fås:

$$\begin{aligned} E(\hat{p}) &= E\left(\frac{X}{n}\right) = \frac{np}{n} = p \\ \text{Var}(\hat{p}) &= \text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2} \text{Var}(X) = \frac{p(1-p)}{n} \end{aligned}$$

Eksempel 1

Venstrehåndede:

p = Andelen af venstrehåndede i Danmark

og/eller:

Kvindelige ingeniørstuderende:

p = Andelen af kvindelige ingeniørstuderende

Example 1

Venstrehåndede (observeret: $x = 10$ ud af $n = 100$):

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{10/100(1-10/100)}{100}} = 0.03$$

$$0.10 \pm 1.96 \cdot 0.03 = 0.10 \pm 0.059 = [0.041, 0.159]$$

Bedre “small sample”-metode – “plus 2”-tilgangen (Remark 7.7)

Anvend samme formel på $\tilde{x} = 10 + 2 = 12$ og $\tilde{n} = 100 + 2 + 2 = 104$:

$$\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}} = \sqrt{\frac{12/104(1-12/104)}{104}} = 0.031$$

$$0.115 \pm 1.96 \cdot 0.031 = 0.115 \pm 0.061 = [0.054, 0.177]$$

Fejlmarginen (Margin of Error (ME))

Fejlmarginen

ved et $(1 - \alpha)100\%$ konfidensniveau er:

$$ME = z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

hvor vi estimerer p med $\hat{p} = \frac{x}{n}$.

Fejlmarginen:

- svarer til den halve bredde af $(1 - \alpha)100\%$ konfidensintervallet.
- Beskriver den “mindst ønskede præcision” på estimatet \hat{p} .

Præcision og stikprøvestørrelse

Forsøgsplanlægning:

Hvor stor skal stikprøvestørrelsen være for at opnå en given præcision?

Metode 7.13

Ønskes en given fejlmargen, ME, med $(1 - \alpha)100\%$ konfidens, da kræves følgende stikprøvestørrelse:

$$n = p(1-p) \left(\frac{z_{1-\alpha/2}}{ME} \right)^2$$

Præcision og stikprøvestørrelse

Method 7.13

Ønskes en given fejlmargen, ME, med $(1 - \alpha)100\%$ konfidens, men hvor vi *ikke* har et fornuftigt gæt på p , da kræves følgende stikprøvestørrelse:

$$n = \frac{1}{4} \left(\frac{z_{1-\alpha/2}}{ME} \right)^2$$

thi "worst case" er $p = \frac{1}{2}$.

Overview

- 1 Introduktion
- 2 Konfidensinterval for én andel
 - Stikprøvestørrelse, forsøgsplanlægning
- 3 Hypotesetest for én andel
- 4 Konfindensinterval og hypotesetest for to andele
- 5 Hypotesetest for flere andele
- 6 Statistik for antalstabeller

Eksempel 1 – fortsat

Venstrehåndethed:

Antag at vi ønsker $ME = 0.01$ (hvor $\alpha = 0.05$) – hvad skal n så være?

Antag $p \approx 0.10$:

$$n = 0.1 \cdot 0.9 \left(\frac{1.96}{0.01} \right)^2 = 3467.4 \approx 3468$$

Uden antagelse om hvad p er:

$$n = \frac{1}{4} \left(\frac{1.96}{0.01} \right)^2 = 9604$$

Trin i et hypotesetest – overblik (repetition!)

- 1 Opstil hypotesen og vælg signifikansniveau α
- 2 Beregn teststørrelse
- 3 Beregn p -værdi ud fra teststørrelsen og den relevante fordeling. Sammenlign p -værdien med signifikansniveauet α og konkluder.
- 4 (Alternativt: lav konklusion ud fra kritiske værdier).

Hypotesetest for én andel

Vi betragter en nul- og alternativ hypotese for én andel p :

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0$$

Som sædvanligt enten accepteres H_0 eller afvises H_0 .

Hypotesetest: teststørrelsen

Theorem 7.10 og Method 7.11

Hvis stikprøven er tilstrækkelig stor ($np_0 > 15$ og $n(1 - p_0) > 15$), bruges teststørrelsen:

$$z_{\text{obs}} = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}}$$

Under nulhypotesen gælder at den stokastiske variabel Z (tilnærmelsesvis) følger en standard normalfordeling, dvs. $Z \sim N(0, 1^2)$.

Hypotesetest: p -værdi og konklusion (Method 7.11)

Find p -værdien (evidens imod nulhypotesen):

- $2P(Z > |z_{\text{obs}}|)$

Test ved brug af kritisk værdi:

Vi afviser nulhypotesen hvis $z_{\text{obs}} < -z_{1-\alpha/2}$ eller $z_{\text{obs}} > z_{1-\alpha/2}$.

Eksempel 1 – fortsat

Er halvdelen af alle danskere venstrehådede?

$$H_0 : p = 0.5, H_1 : p \neq 0.5$$

Teststørrelse:

$$z_{\text{obs}} = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}} = \frac{10 - 100 \cdot 0.5}{\sqrt{100 \cdot 0.5(1 - 0.5)}} = -8$$

p -værdi:

$$2 \cdot P(Z > 8) = 1.2 \cdot 10^{-15}$$

Der er meget stærk evidens mod nulhypotesen.

Eksempel 1 – fortsat

Hypotesetest in R

```
prop.test(10, 100, p = 0.5, correct = FALSE)
##
## 1-sample proportions test without continuity correction
##
## data: 10 out of 100, null probability 0.5
## X-squared = 64, df = 1, p-value = 1e-15
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.05523 0.17437
## sample estimates:
##      p
## 0.1
```

Overview

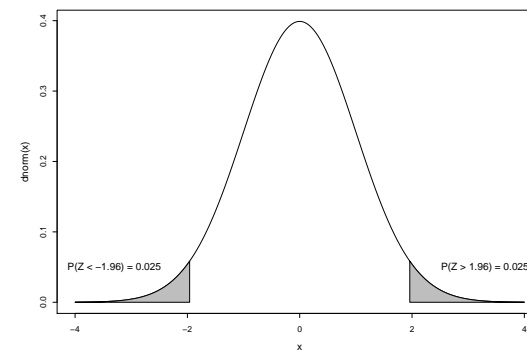
- 1 Introduktion
- 2 Konfidensinterval for én andel
 - Stikprøvestørrelse, forsøgsplanlægning
- 3 Hypotesetest for én andel
- 4 **Konfidensinterval og hypotesetest for to andele**
- 5 Hypotesetest for flere andele
- 6 Statistik for antalstabeller

Eksempel 1 – fortsat

Ved brug af kritiske værdier:

$$z_{0.975} = 1.96$$

Da $z_{\text{obs}} = -8$ er (meget) mindre end -1.96 så afvises hypotesen.



Konfidensinterval for forskellen på to andele

Method 7.15

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2} \cdot \hat{\sigma}_{\hat{p}_1 - \hat{p}_2}$$

hvor

$$\hat{\sigma}_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

Tommelfingerregel

Både $n_i p_i \geq 10$ og $n_i(1 - p_i) \geq 10$ for $i = 1, 2$.

Hypotesetest for forskellen på to andele, Method 7.18

Two sample proportions hypothesis test

Såfremt man ønsker at sammenligne to andele (her vist for et tosidet alternativ)

$$H_0: p_1 = p_2$$

$$H_1: p_1 \neq p_2$$

skal vi bruge teststørrelsen

$$z_{\text{obs}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad \text{hvor } \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Og for tilstrækkeligt store stikprøver:

Brug standardnormalfordelingen igen!

Example 2

I et studie (USA, 1975) undersøgtes sammenhængen mellem p-piller og risikoen for blodprop i hjertet.

	Blodprop	Ikke blodprop
p-piller	23	34
Ikke p-piller	35	132

Estimer i hver stikprøve

$$\hat{p}_1 = \frac{23}{57} = 0.4035, \quad \hat{p}_2 = \frac{35}{167} = 0.2096$$

Fælles estimat:

$$\hat{p} = \frac{23 + 35}{57 + 167} = \frac{58}{224} = 0.2589$$

Eksempel 2

Er der en sammenhæng mellem brugen af p-piller og risikoen for blodprop i hjertet?

I et studie (USA, 1975) undersøgtes sammenhængen mellem p-piller og risikoen for blodprop i hjertet.

	Blodprop	Ikke blodprop
p-piller	23	34
Ikke p-piller	35	132

Undersøg om der er sammenhæng mellem brug af p-piller og risiko for blodprop i hjertet. Anvend signifikansniveau $\alpha = 5\%$.

Eksempel 2 – fortsat

`prop.test`: test om to andele er ens i R

```
# Read data table into R
pill.study <- matrix(c(23, 34, 35, 132),
                    ncol = 2, byrow = TRUE)
colnames(pill.study) <- c("Blood Clot", "No Clot")
rownames(pill.study) <- c("Pill", "No pill")
pill.study

# Test whether probabilities are equal for the two groups
prop.test(pill.study, correct = FALSE)
```

Eksempel 2 – fortsat

prop.test: test om to andele er ens i R

```
##           Blood Clot No Clot
## Pill           23      34
## No pill          35     132
```

```
##
## 2-sample test for equality of proportions without continuity
## correction
##
## data: pill.study
## X-squared = 8.3, df = 1, p-value = 0.004
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.05239 0.33546
## sample estimates:
## prop 1 prop 2
## 0.4035 0.2096
```

Overview

- 1 Introduktion
- 2 Konfidensinterval for én andel
 - Stikprøvestørrelse, forsøgsplanlægning
- 3 Hypotesetest for én andel
- 4 Konfidensinterval og hypotesetest for to andele
- 5 Hypotesetest for flere andele
- 6 Statistik for antalstabeller

Hypotesetest for flere andele

Sammenligning af c andele

I nogle tilfælde kan man være interesseret i at vurdere om to eller flere binomialfordlinger har den samme parameter p , dvs. man er interesseret i at teste nulhypotesen:

$$H_0: p_1 = p_2 = \dots = p_c = p$$

mod den alternative hypotese om at disse andele ikke er ens.

Hypotesetest for flere andele

Tablet af observerede antal for c stikprøver:

	Sample 1	Sample 2	...	Sample c	Total
Succes	x_1	x_2	...	x_c	x
Failure	$n_1 - x_1$	$n_2 - x_2$...	$n_c - x_c$	$n - x$
Total	n_1	n_2	...	n_c	n

Fælles (gennemsnitligt) estimat:

Under nulhypotesen er estimatet for p :

$$\hat{p} = \frac{x}{n}$$

Hypotesetest for flere andele

Fælles (gennemsnitligt) estimat:

Under nulhypotesen er estimatet for p :

$$\hat{p} = \frac{x}{n}$$

”Brug” dette fælles estimat i hver gruppe:

Hvis nulhypotesene er sand, så forventer vi at den j 'te gruppe har e_{1j} succeser og e_{2j} fiaskoer, hvor

$$e_{1j} = n_j \cdot \hat{p} = \frac{n_j \cdot x}{n}$$

$$e_{2j} = n_j(1 - \hat{p}) = \frac{n_j \cdot (n - x)}{n}$$

Hypotesetest for flere andele

Make a table with the *expected* counts for the c samples:

e_{ij}	Sample 1	Sample 2	...	Sample c	Total
Success	e_{11}	e_{12}	...	e_{1c}	x
Failure	e_{21}	e_{22}	...	e_{2c}	$n - x$
Total	n_1	n_2	...	n_c	n

General formel for beregning af forventede værdier i antalstabeller:

$$e_{ij} = \frac{(i\text{'th row total}) \cdot (j\text{'th column total})}{\text{total}}$$

Beregning af teststørrelsen - Method 7.20

Teststørrelsen bliver

$$\chi_{\text{obs}}^2 = \sum_{i=1}^2 \sum_{j=1}^c \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

hvor o_{ij} er det *observerede* antal i celle (i, j) og e_{ij} er det *forventede* antal i celle (i, j) .

Find p -værdien eller brug kritisk værdi – Method 7.20

Stikprøvefordeling for teststørrelse (under H_0):
 χ^2 -fordeling med $(c - 1)$ frihedsgrader (approx.)

Kritisk værdi metode:

Hvis $\chi_{\text{obs}}^2 > \chi_{\alpha}^2(c - 1)$, så afvises nulhypotesen.

Tommelfingerregel for om testen er valid:

Alle forventede værdier $e_{ij} \geq 5$.

Eksempel 2 – fortsat

De observerede værdier o_{ij}

Observed	Blodprop	Ikke blodprop
p-piller	23	34
Ikke p-piller	35	132

Eksempel 2 – fortsat

Beregn de forventede værdier e_{ij}

Expected	Blodprop	Ikke blodprop	Total
p-piller			57
Ikke p-piller			167
Total	58	166	224

Eksempel 2 – fortsat

Brug "reglen" for forventede værdier fire gange, e.g.:

$$e_{22} = \frac{167 \cdot 166}{224} = 123.76$$

Eksempel 2 – fortsat

Teststørrelsen (husk at inkludere alle celler):

$$\chi_{\text{obs}}^2 = \frac{(23 - 14.76)^2}{14.76} + \frac{(34 - 42.24)^2}{42.24} + \frac{(35 - 43.24)^2}{43.24} + \frac{(132 - 123.76)^2}{123.76} = 8.33$$

De forventede værdier e_{ij} :

Expected	Blodprop	Ikke blodprop	Total
p-piller	14.76	42.24	57
Ikke p-piller	43.24	123.76	167
Total	58	166	224

Kritisk værdi:

`qchisq(0.95, 1)`

[1] 3.841

Konklusion:

Vi afviser nulhypotesen – der er altså en signifikant højere risiko for blodpropper i gruppen, som indtager p-piller.

Eksempel 2 – fortsat

`chisq.test` for at teste om to forhold er ens i R.

```
# Test whether probabilities are equal for the two groups
chisq.test(pill.study, correct = FALSE)

##
## Pearson's Chi-squared test
##
## data: pill.study
## X-squared = 8.3, df = 1, p-value = 0.004

# Expected values
chisq.test(pill.study, correct = FALSE)$expected

##          Blood Clot No Clot
## Pill      14.76    42.24
## No pill    43.24   123.76
```

Overview

- 1 Introduktion
- 2 Konfidensinterval for én andel
 - Stikprøvestørrelse, forsøgsplanlægning
- 3 Hypotesetest for én andel
- 4 Konfindensinterval og hypotesetest for to andele
- 5 Hypotesetest for flere andele
- 6 Statistik for antalstabeller

Eksempel 3: Analyse af en antalstabel

En 3×3 -tabel – 3 stikprøver, 3-kategori udfald

	4 weeks bef	2 weeks bef	1 week bef
Candidate I	79	91	93
Candidate II	84	66	60
Undecided	37	43	47
	$n_1 = 200$	$n_2 = 200$	$n_3 = 200$

Er stemmefordelingen ens?

$$H_0: p_{i1} = p_{i2} = p_{i3}, \quad i = 1, 2, 3.$$

En anden slags antalstabel

En 3×3 -tabel – 1 stikprøve, to 3-kategori variable:

	bad	average	good
bad	23	60	29
average	28	79	60
good	9	49	63

Er der uafhængighed mellem inddelingskriterier?

$$H_0: p_{ij} = p_i \cdot p_j$$

Beregning af teststørrelse – uanset typen af tabel 7.22

I en antalstabel med r rækker og c søjler, da er teststørrelsen:

$$\chi_{\text{obs}}^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

hvor o_{ij} er observeret antal i celle (i, j) , og e_{ij} er det *forventede antal* i celle (i, j) (under hypotesen).

Generel formel for beregning af forventede værdier i antalstabeller:

$$e_{ij} = \frac{(i\text{'th row total}) \cdot (j\text{'th column total})}{\text{total}}$$

Find p -værdi eller brug kritisk værdi - Method 7.22

Stikprøvefordeling for test-størrelse unde H_0 :

χ^2 -fordeling med $(r - 1)(c - 1)$ frihedsgrader.

Kritisk værdi-metode:

Såfremt $\chi_{\text{obs}}^2 > \chi_{\alpha}^2$ med $(r - 1)(c - 1)$ frihedsgrader, da forkastes nulhypotesen.

Tommelfingerregel for validitet af test:

Alle forventede værdier $e_{ij} \geq 5$.

Eksempel 3 – fortsat

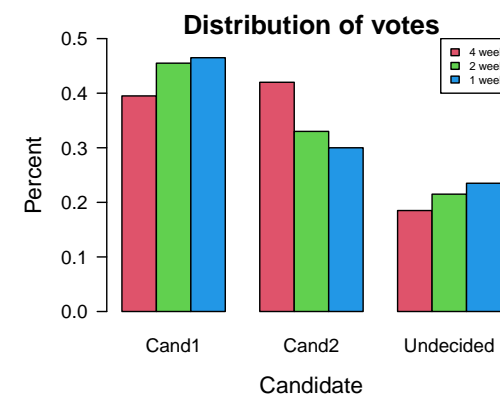
`chisq.test` for antalstabeller

```
# Read data table into R
poll <-matrix(c(79, 91, 93, 84, 66, 60, 37, 43, 47),
             ncol = 3, byrow = TRUE)
colnames(poll) <- c("4 weeks", "2 weeks", "1 week")
rownames(poll) <- c("Cand1", "Cand2", "Undecided")

# Show column percentages
prop.table(poll, 2)

##           4 weeks 2 weeks 1 week
## Cand1      0.395  0.455  0.465
## Cand2      0.420  0.330  0.300
## Undecided  0.185  0.215  0.235
```

Eksempel 3 – fortsat



Eksempel 3 – fortsat

```

# Testing for same distribution in the three populations
chisq.test(poll, correct = FALSE)

##
## Pearson's Chi-squared test
##
## data:  poll
## X-squared = 7, df = 4, p-value = 0.1

# Expected values
chisq.test(poll, correct = FALSE)$expected

##           4 weeks 2 weeks 1 week
## Cand1      87.67   87.67  87.67
## Cand2      70.00   70.00  70.00
## Undecided  42.33   42.33  42.33

```

Overview

- 1 Introduktion
- 2 Konfidensinterval for én andel
 - Stikprøvestørrelse, forsøgsplanlægning
- 3 Hypotesetest for én andel
- 4 Konfidensinterval og hypotesetest for to andele
- 5 Hypotesetest for flere andele
- 6 Statistik for antalstabeller