Kursus 02402/02323 Introducerende Statistik

Forelæsning 12: Inferens for andele

Klaus K. Andersen og Per Bruun Brockhoff

DTU Compute, Statistik og Dataanalyse
Danmarks Tekniske Universitet
2800 Lyngby – Danmark

e-mail: klaus@cancer.dk

Oversigt

1. Intro
2. Konfidensinterval for én andel
   - Eksempel 1
   - Bestemmelse af stikprøvestørrelse
     - Eksempel 1 - fortsat
3. Hypotesetest for én andel
   - Eksempel 1 - fortsat
4. Konfidensinterval og hypotesetest for to andele
   - Eksempel 2
5. Hypotesetest for flere andele
   - Eksempel 2 - fortsat
6. Analyse af antalstabeller
7. R

Forskellige analyse/data-situationer

Gennemsnit for kvantitative data:

- Hypotesetest/KI for én middelværdi (one-sample)
- Hypotesetest/KI for to middelværdier (two samples)
- Hypotesetest/KI for flere middelværdier (K samples)

I dag: Andele:

- Hypotesetest/KI for én andel
- Hypotesetest/KI for to andele
- Hypotesetest for flere andele
- Hypotesetest for flere "multi-categorical" andele

Estimation af andele

- Estimation af andele fås ved at observere antal gange x en hændelse har indtruffet ud af n forsøg:

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

$$\hat{p} \in [0; 1]$$

Konfidensinterval for én andel

Method 7.3

Såfremt der haves en stor stikprøve, fås et $(1-\alpha)\%$ konfidensinterval for p

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

Hvordan?

Følger af at approximere binomialfordelingen med normalfordelingen.

As a rule of thumb

the normal distribution gives a good approximation of the binomial distrinution if np and $n(1-p)$ are both greater than 15

Konfidensinterval for én andel

Middelværdi og varians i binomialfordelingen, eNote2:

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

This means that

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

Eksempel 1

Venstrehåndede:

$p =$ Andelen af venstrehåndede i Danmark

og/eller:

Kvindelige ingeniørstuderende:

$p =$ Andelen af kvindelige ingeniørstuderende

Eksempel 1

Venstrehåndede:

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

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

Bedre "small sample" metode - "plus 2-approach":(Remark 7.7)

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

$$\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}} = \sqrt{\frac{12/104(1-12/104)}{104}} = 0.031328$$

$$0.1154 \pm 1.96 \cdot 0.03132 \Leftrightarrow 0.1154 \pm 0.0614 \Leftrightarrow [0.054, 0.177]$$

"Margin of Error" på estimat

Margin of Error

med $(1-\alpha)$% konfidens bliver

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

hvor et estimat af p fås ved $p = \frac{x}{n}$

Bestemmelse af stikprøvestørrelse

Method 7.12

Såfremt man højst vil tillade en Margin of Error ME med $(1-\alpha)$% konfidens, bestemmes den nødvendige stikprøvestørrelse ved

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

Bestemmelse af stikprøvestørrelse

Method 7.12

Såfremt man højst vil tillade en Margin of Error ME med $(1-\alpha)$% konfidens, og p ikke kendes, bestemmes den nødvendige stikprøvestørrelse ved

$$n = \frac{1}{4}[\frac{z_{1-\alpha/2}}{ME}]^2$$

idet man får den mest konservative stikprøvestørrelse ved at vælge $p = \frac{1}{2}$

Eksempel 1 - fortsat

Venstrehåndede:

Antag vi ønsker $ME = 0.01$ (med $\alpha = 0.05$) - hvad skal n 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 størrelsen af p:

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

Trin ved Hypoteseprøvning

1. Opstil hypoteser og vælg signifikansniveau α
2. Beregn teststørrelse
3. Beregn p-værdi (eller kritisk værdi)
4. Fortolk p-værdi og/eller Sammenlign p-værdi og signifikansniveau og drag en konklusion

(Alternativ 4. Sammenlign teststørrelse og kritisk værdi og drag en konklusion)

Hypotesetest for én andel

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

$$H_0: \quad p = p_0$$
$$H_1: \quad p \neq p_0$$

Man vælger som sædvanligt enten at acceptere H_0 eller at forkaste H_0

Beregning af teststørrelse

Theorem 7.9 og Method 7.10

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

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

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

Test ved brug af p-værdi (Method 7.10)

Find p-værdien (evidence mod nulhypotesen):

- If two-sided: $2P(Z > |z_{\text{obs}}|)$
- If one-sided "less": $P(Z < z_{\text{obs}})$
- If one-sided "greater": $P(Z > z_{\text{obs}})$

Test ved brug af kritisk værdi (Method 7.10)

Afhængig af den alternative hypotese fås følgende kritiske værdier

Alternativ hypotese	Afvis nul-hypotese hvis
$p < p_0$	$z_{\text{obs}} < -z_{1-\alpha}$
$p > p_0$	$z_{\text{obs}} > z_{1-\alpha}$
$p \neq p_0$	$z_{\text{obs}} < -z_{1-\alpha/2}$ eller $z_{\text{obs}} > z_{1-\alpha/2}$

Eksempel 1 - fortsat

Er halvdelen af alle danskere venstrehåndede?

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

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:

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

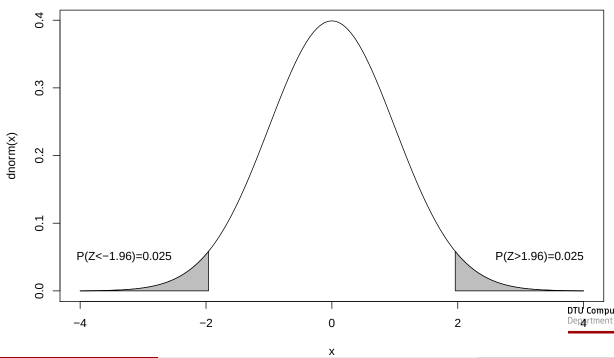
Der er meget stærk evidence imod nulhypotesen - vi kan forkaste denne (med $\alpha = 0.05$).

Eksempel 1 - fortsat

Evt med kritisk værdi i stedet:

$$z_{0.975} = 1.96$$

Idet $z_{\text{obs}} = -8$ er (meget) mindre end -1.96 kan vi forkaste hypotesen.



Konfidensinterval for to andele

Method 7.14

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

hvor

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

Rule of thumb:

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

Hypotesetest for to andele, Method 7.17

Two sample proportions hypothesis test

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

$$H_0: \quad p_1 = p_2$$
$$H_1: \quad p_1 \neq p_2$$

Fås teststørrelsen:

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

Og for passende store stikprøver:

Brug standardnormalfordelingen igen.

Eksempel 2

Sammenhæng mellem brug af p-piller og risikoen for hjerteinfarkt

I et studie (USA, 1975) undersøgte man dette. Fra et hospital havde man indsamlet følgende stikprøve

	Infarkt	Ikke infarkt
p-piller	23	34
Ikke p-piller	35	132

Er der sammenhæng mellem brug af p-piller og sygdomsrisiko

Udfør et test for om der er sammenhæng mellem brug af p-piller og risiko for hjerteinfarkt. Anvend signifikansniveau $\alpha = 5\%$

Eksempel 2

Sammenhæng mellem brug af p-piller og risikoen for hjerteinfarkt

	Infarkt	Ikke infarkt	Total
p-piller	23	34	$n_1 = 57$
Ikke p-piller	35	132	$n_2 = 167$
	$x = 58$		$n = 224$

Estimater i hver stikprøve

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

Fælles estimat:

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

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 nul-hypotesen

$$H_0: \quad p_1 = p_2 = ... = p_c = p$$

mod en alternativ hypotese at disse andele ikke er ens

Hypotesetest for flere andele

Tabel af observerede antal for k stikprøver:

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

Fælles (gennemsnitlig) estimat:

Under nul-hypotesen fås et estimat for p:

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

Hypotesetest for flere andele

Fælles (gennemsnitlig) estimat:

Under nul-hypotesen fås et estimat for p:

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

"Brug" dette fælles estimat i hver gruppe:

såfremt nul-hypotesen gælder, vil vi forvente at den j'te gruppe har e_{1j} successer 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

Generel formel for beregning af forventede værdier i antalstabeller:

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

Beregning af teststørrelse - Method 7.19

Teststørrelsen bliver

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

hvor o_{ij} er observeret antal i celle (i, j) og e_{ij} er forventet antal i celle (i, j)

Find p-værdi eller brug kritisk værdi - Method 7.19

Stikprøvefordeling for test-størrelse:

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

Kritisk værdi metode

Såfremt $\chi^2_{\text{obs}} > \chi^2_{\alpha}(c-1)$ forkastes nul-hypotesen

Rule of thumb for validity of the test:

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

Eksempel 2 - fortsat

De OBSERVEREDE værdier o_{ij}

Observerede	Infarkt	Ikke infarkt
p-piller	23	34
Ikke p-piller	35	132

Eksempel 2 - fortsat

Beregn de FORVENTEDE værdier e_{ij}

Forventede	Infarkt	Ikke infarkt	Total
p-piller			57
Ikke p-piller			167
Total	58	166	224

Eksempel 2 - fortsat

Brug "reglen" for forventede værdier fire gange, f.eks. :

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

De FORVENTEDE værdier e_{ij}

Forventede	Infarkt	Ikke infarkt	Total
p-piller	14.76	42.24	57
Ikke p-piller	43.24	123.76	167
Total	58	166	224

Eksempel 2 - fortsat

Teststørrelsen:

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

Kritisk værdi:

[1] 3.8415

Konklusion:

Vi forkaster hulhypotesen - der ER en signifikant forhøjet sygdomsrisiko i p-pille gruppen.

Analyse af antalstabeller

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

	4 uger før	2 uger før	1 uge før
Kandidat I	79	91	93
Kandidat II	84	66	60
ved ikke	37	43	47
	$n_1 = 200$	$n_2 = 200$	$n_3 = 200$

Er stemmefordelingen ens?

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

Analyse af antalstabeller

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

	dårlig	middel	god
dårlig	23	60	29
middel	28	79	60
god	9	49	63

Er der uafhængighed mellem inddelingskriterier?

$$H_0: \ p_{ij} = p_{i\cdot} p_{\cdot j}$$

Beregning af teststørrelse – uanset type af tabel

I en antalstable med r rækker og c søjler, fås teststørrelsen

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

hvor o_{ij} er observeret antal i celle (i, j) og e_{ij} er forventet antal i celle (i, j)

Generel formel for beregning af forventede værdier i antalstabeller:

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

Find p-værdi eller brug kritisk værdi - Method 7.21

Stikprøvefordeling for test-størrelse:

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

Kritisk værdi metode

Såfremt $\chi^2_{\text{obs}} > \chi^2_\alpha$ med $(r-1)(c-1)$ frihedsgrader forkastes nul-hypotesen

Rule of thumb for validity of the test:

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

R: prop.test - een andel

```
# WITHOUT CONTINUITY CORRECTIONS

prop.test(518, 1154, p = 0.5, correct = FALSE)
```


R: prop.test - to andele

```
colnames(pill.study) <- c("Blood Clot", "No Clot")
rownames(pill.study) <- c("Pill", "No pill")

# TESTING THAT THE PROBABILITIES FOR THE TWO GROUPS ARE EQUAL
prop.test(pill.study, correct = FALSE)
```


R: chisq.test - to andele

```
#IF WE WANT THE EXPECTED NUMBERS SAVE THE TEST IN AN OBJECT
chi <- chisq.test(pill.study, correct = FALSE)
#THE EXPECTED VALUES
chi$expected
```

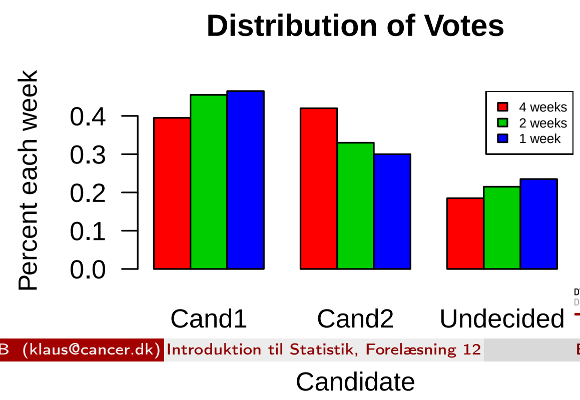
R: chisq.test - antalstabeller

```
colnames(poll) <- c("4 weeks", "2 weeks", "1 week")
rownames(poll) <- c("Cand1", "Cand2", "Undecided")

#COLUMN PERCENTAGES
colpercent<-prop.table(poll, 2)
colpercent
```


R: chisq.test - antalstabeller

```
barplot(t(colpercent), beside = TRUE, col = 2:4, las = 1,
        ylab = "Percent each week", xlab = "Candidate",
        main = "Distribution of Votes")
legend( legend = colnames(poll), fill = 2:4,"topright", cex = 0.5)
par(mar=c(5,4,4,2)+0.1)
```



R: chisq.test - antalstabeller

```
chi

#EXPECTED VALUES
chi$expected
```


Oversigt

1. Intro
2. Konfidensinterval for én andel
   - Eksempel 1
   - Bestemmelse af stikprøvestørrelse
     - Eksempel 1 - fortsat
3. Hypotesetest for én andel
   - Eksempel 1 - fortsat
4. Konfidensinterval og hypotesetest for to andele
   - Eksempel 2
5. Hypotesetest for flere andele
   - Eksempel 2 - fortsat
6. Analyse af antalstabeller
7. R