Kursus 02403 Introduktion til Matematisk Statistik

Forelæsning 10: Inferens for andele

Jan Kloppenborg Møller
DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 016
Danmarks Tekniske Universitet
2800 Lyngby – Danmark

e-mail: jkmo@dtu.dk

Oversigt

Oversigt

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:

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

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

Oversigt

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

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{np}{n} = p$$

$$Var(\hat{p}) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2} Var(X) = \frac{p(1-p)}{n}$$

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

Eksempler

Venstrehåndede:

$$p = \text{Andelen af venstrehåndede i Danmark}$$

og/eller:

Kvindelige ingeniørstuderende:

$$p = \text{Andelen af kvindelige ingeniørstuderende}$$

Eksempel

- Find et KI for andelen af stemmer på rød blok
- Find et KI for andelen af stemmer på Moderaterne

"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 7.13

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

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

Bestemmelse af stikprøvestørrelse

Method 7.13

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

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

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

Eksempel

Hvad er stikprøvestørrelsen hvis man ønsker en Magin of error på maksimalt den målte afstand mellem rød blok og $p = 0.5$?

Oversigt

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 7.10 og Method 7.11

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 stokastiske variabel Z følger en standard normalfordeling, dvs. $Z \sim N(0, 1^2)$

Find p-værdien for to-sidet alternativ (evidence mod nulhypotesen):

- If two-sided: $2P(Z > |z_{\text{obs}}|)$

Kan også gøres ved brug af kritisk værdi.

Eksempel - Hypotesetest

Udfør hypotetesten

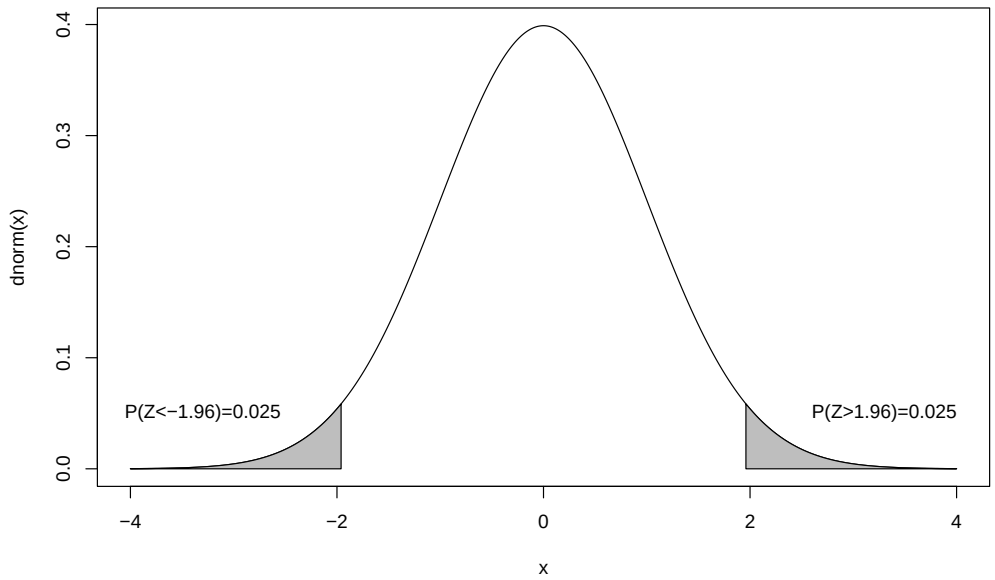
$$H_0 : \text{Der er dødt løb mellem blokkene}$$

mod et rimeligt alternativ.

Eksempel

Evt med kritisk værdi i stedet:

$$z_{0.975} = 1.96$$



Oversigt

Konfidensinterval for 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

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

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 7.18

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 - Konfidens interval og hypotese test for 2 andele

- Er der forskel på fordelingen mellem blokke i April og Marts (brug KI)
- Er der forskel på Alternativets stemmeandel (brug hypotese test)?

Oversigt

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 7.20

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 7.20

Stikprøvefordeling for test-størrelse:

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

Kritisk værdi metode

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

Rule of thumb for validity of the test:

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

Eksempel

- Er fordelingen mellem blokke ens i de opgivne meningsmålinger?

Oversigt

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 7.22

Stikprøvefordeling for test-størrelse:

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

Kritisk værdi metode

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

Rule of thumb for validity of the test:

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

Eksempel

- Er der en tidslig udvikling i menningsmålingerne?

Oversigt

R: prop.test - een andel

```
# TESTING THE PROBABILITY = 0.5 WITH A TWO-SIDED ALTERNATIVE
# WE HAVE OBSERVED 518 OUT OF 1154
# WITHOUT CONTINUITY CORRECTIONS

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


##
##  1-sample proportions test without continuity correction
##
## data:  518 out of 1154, null probability 0.5
## X-squared = 12.1, df = 1, p-value = 0.00051
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.42039 0.47769
## sample estimates:
##       p
## 0.44887
```

R: prop.test - to andele

```
#READING THE 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")

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


##
##  2-sample test for equality of proportions without continuity correction
##
## data:  pill.study
## X-squared = 8.33, df = 1, p-value = 0.0039
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.052395 0.335461
## sample estimates:
##  prop 1  prop 2
## 0.40351 0.20958
```

R: chisq.test - to andele

```
# CHI2 TEST FOR TESTING THE PROBABILITIES FOR THE TWO GROUPS ARE EQUAL
chisq.test(pill.study, correct = FALSE)

##
##  Pearson's Chi-squared test
##
## data:  pill.study
## X-squared = 8.33, df = 1, p-value = 0.0039

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

##          Blood Clot No Clot
## Pill         14.759  42.241
## No pill      43.241 123.759
```

R: chisq.test - antalstabeller

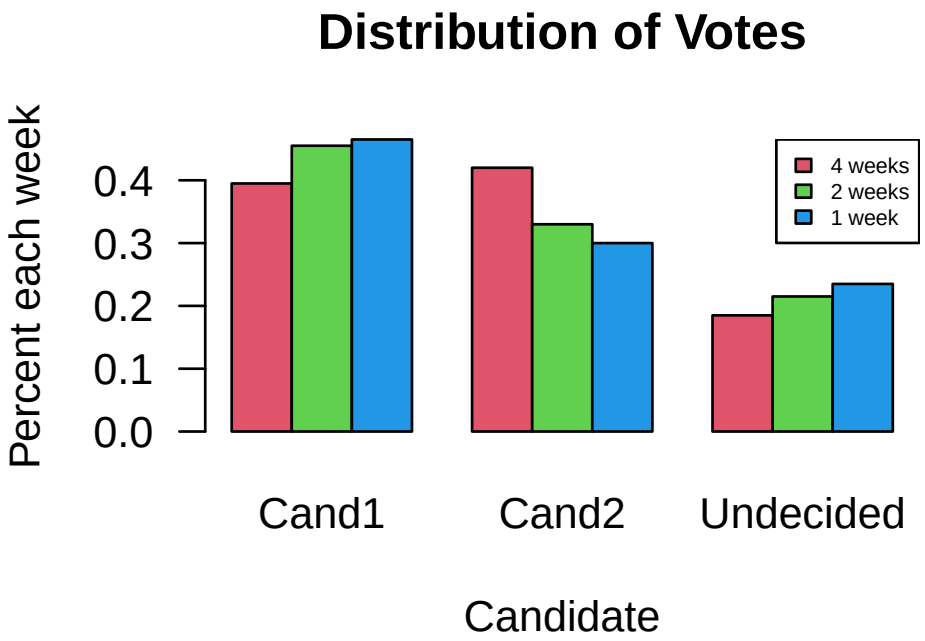
```
#READING THE TABLE INTO  R
poll <-matrix(c(79, 91, 93, 84, 66, 60, 37, 43, 47), ncol = 3, byrow = TRUE)
colnames(poll) <- c("4 weeks", "2 weeks", "1 week")
rownames(poll) <- c("Cand1", "Cand2", "Undecided")

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

```
##           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
```

R: chisq.test - antalstabeller

```
# Plotting percentages
par(mar=c(5,4,4.1,2)+0.1)
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

```
#TESTING SAME DISTRIBUTION IN THE THREE POPULATIONS
chi <- chisq.test(poll, correct = FALSE)
chi

##
##  Pearson's Chi-squared test
##
## data:  poll
## X-squared = 6.96, df = 4, p-value = 0.14

#EXPECTED VALUES
chi$expected

##           4 weeks 2 weeks 1 week
## Cand1      87.667  87.667 87.667
## Cand2      70.000  70.000 70.000
## Undecided  42.333  42.333 42.333
```

Oversigt