02323 Introduktion til statistik

Uge 10: Kategorisk data og andele

Nicolai Siim Larsen
DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Dagsorden

Dagsorden

Forskellige analyser og data i 02323

Middelværdi for kvantitative data

- Hypotesetest/KI for én middelværdi baseret på én stikprøve
- Hypotesetest/KI for to middelværdier baseret på to stikprøver
- Hypotesetest/KI for flere middelværdier baseret på K stikprøver

I dag: Andele/Forholdstal (proportions)

- Hypotesetest/KI for én andel baseret på én stikprøve
- Hypotetseest/KI for to andele baseret på to stikprøver
- Hypotesetest for flere andele baseret på c stikprøver
- Hypotesetest for multikategoriske data

Estimation af andele

- Estimation af en andel (sandsynlighed) fås ved at observere antallet af gange en hændelse har indtruffet (x) i 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.

Dagsorden

Konfidensinterval for én andel

Metode 7.3

Hvis stikprøven er stor, så er $(1-\alpha)$-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 $X \sim \text{bin}(n,p)$. Normalfordelingen er en god tilnærmelse til binomialfordelingen hvis np og $n(1-p)$ (forventede antal succeser og fiaskoer) begge er mindst 15.

Konfidensinterval for én andel

Middelværdi og varians i binomialfordelingen, sektion 2.21

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

Altså fås:

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

Eksempel på andele

Venstrehåndede:

p : Andelen af venstrehåndede i Danmark

eller:

Kvindelige ingeniørstuderende:

p : Andelen af kvindelige ingeniørstuderende

Eksempel 1: Et 95%-konfidensinterval

Venstrehåndede (observeret data er $x = 10$ 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 metode til små stikprøver: "plus 2"-tilgangen (Bemærkning 7.7)

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

$$\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}} = \sqrt{\frac{12/104(1-12/104)}{104}} = 0.03$$

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

Fejlmarginen (ME: Margin of Error)

Fejlmarginen

ved et $(1-\alpha)$-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 $(1-\alpha)$-konfidensintervallet.
- Beskriver den forventede præcision (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 forventet (given) fejlmargin (ME) i et $(1-\alpha)$-konfidensinterval, kræves følgende stikprøvestørrelse:

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

hvor p er et fornuftigt gæt.

Præcision og stikprøvestørrelse

Metode 7.13

Ønskes en forventet (given) fejlmargin (ME) i et $(1-\alpha)$-konfidensinterval, men hvor vi *ikke* har et fornuftigt gæt på p, da kræves følgende stikprøvestørrelse:

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

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

Eksempel 1 – fortsat

Venstrehåndede:

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

Baseret på stikprøven gættes på $p = 0.10$:

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

Uden antagelse/gæt om hvad p er:

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

Dagsorden

Trin i en hypotesetest – Overblik (repetition!)

1. Opstil nulhypotesen og vælg et signifikansniveau α
2. Beregn den observerede teststørrelse
3. Beregn p-værdien ud fra den observerede teststørrelse og den relevante fordeling
4. Sammenlign p-værdien med signifikansniveauet α og konkludér

Alternativt: Sammenlign den observerede teststørrelse med kritiske værdier og konkludér.

Hypotesetest for én andel

Vi betragter en nul- og modhypotese for én andel p og vælger et signifikansniveau α:

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

Som sædvanligt afvises H_0 eller accepteres H_0.

Hypotesetest: Teststørrelsen

Sætning 7.10 og metode 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 teststørrelsen (tilnærmelsesvis) følger en standardnormalfordeling.

Hypotesetest: p-værdi og konklusion (Metode 7.11)

Find p-værdien (evidens imod nulhypotesen):

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

Test ved brug af kritiske værdier:

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 venstrehåndede?

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

Teststørrelsen:

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

p-værdi:

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

Der er meget stærk evidens mod nulhypotesen.

Eksempel 1 – fortsat

Hypotesetest i R

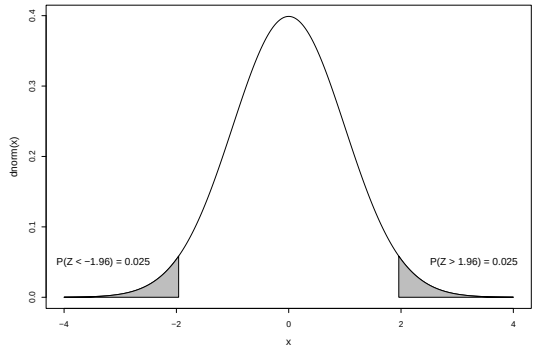
```
prop.test(10, 100, p = 0.5, correct = FALSE)

##
##  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
```


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 nulhypotesen.



Dagsorden

Konfidensinterval for forskellen på to andele

Metode 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 - Metode 7.18

Hypotesetest for to andele

Såfremt man ønsker at sammenligne to andele (her vist for en tosidet modhypotese)

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

skal man bruge teststørrelsen

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

Eksempel 2

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

I et studie (USA, 1975) undersøgtes sammenhængen mellem p-piller og risikoen for blodpropper 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 blodpropper i hjertet. Anvend signifikansniveauet $\alpha = 5\%$.

Eksempel 2

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

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

Estimater i hver stikprøve

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

Fælles estimat:

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

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
```


Dagsorden

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:

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

mod den alternative hypotese om at disse andele ikke er ens (dvs. mindst én er anderledes).

Hypotesetest for flere andele

Tabel af observerede antal for c 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 (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} 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

Tabel med det *forventede* antal i de c stikprøver:

e_{ij}	Stikprøve 1	Stikprøve 2	...	Stikprøve c	Total
Succes	e_{11}	e_{12}	...	e_{1c}	x
Fiasko	e_{21}	e_{22}	...	e_{2c}	$n-x$
Total	n_1	n_2	...	n_c	n

Generel formel for beregning af forventede værdier i antalstabeller:

$$e_{ij} = \frac{(\text{Rækketotal } i) \cdot (\text{Kolonnetotal } j)}{\text{total}}$$

Beregning af teststørrelsen - Metode 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 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 – Metode 7.20

Stikprøvefordeling for teststørrelsen (under H_0):

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

Metode med kritiske værdier:

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

Tommelfingerregel for om testen er valid:

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

Eksempel 2 – fortsat

De *observerede* værdier o_{ij}

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

Eksempel 2 – fortsat

Beregn de *forventede* værdier e_{ij}

Forventede	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, dvs.:

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

De *forventede* værdier e_{ij}:

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

Eksempel 2 – fortsat

Teststørrelsen (husk at inkludere alle celler):

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

Den kritiske værdi:

```
qchisq(0.95, 1)
```

```
[1] 3.841
```

Konklusion:

Vi afviser nulhypotesen. Der er altså en signifikant sammenhæng mellem risikoen for blodpropper i hjertet og brugen af 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
```


Dagsorden

Eksempel 3: Analyse af en antalstabel

En 3×3-tabel: 3 stikprøver med 3 kategoriske udfald

	4 weeks	2 weeks	1 week
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?

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

En anden slags antalstabel

En 3×3-tabel: 1 stikprøve med to variable med 3 kategoriske udfald:

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

Er der uafhængighed mellem inddelingskriterier?

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

Teststørrelsen – uanset typen af tabel: Metode 7.22

I en antalstabel med r rækker og c søjler, da er 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 det observerede antal i celle (i,j), og e_{ij} er det *forventede antal* i celle (i,j) (under nulhypotesen).

Generel formel for beregning af forventede værdier i antalstabeller:

$$e_{ij} = \frac{(\text{Rækketotal } i) \cdot (\text{Kolonnetotal } j)}{\text{total}}$$

Find p-værdi eller brug den kritiske værdi - Metode 7.22

Stikprøvefordeling for teststørrelsen under H_0:

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

Metode med den kritiske værdi:

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

Tommelfingerregel for validitet af test:

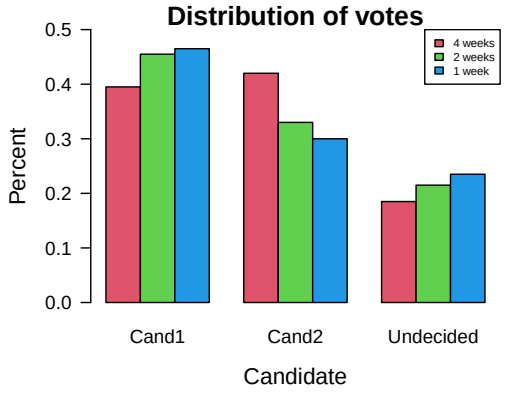
Alle forventede værdier $e_{ij} \geq 5$.

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