Introduktion til Statistik

Forelæsning 12: Inferens for andele

Peder Bacher

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 009
Danmarks Tekniske Universitet
2800 Lyngby – Danmark
e-mail: pbac@dtu.dk

Forår 2017

Kapitel 7: Inferens for andele

Engelsk

Statistik for andele:

- Andel: $p = \frac{x}{n}$ *(x successer ud af n observationer)*
- Specifikke metoder, én, to og $k > 2$ grupper
  - Binær/kategorisk respons

Specifikke metoder:

- Estimation og konfidensintervaller for andele
  - Metoder til store stikprøver vs. til små stikprøver
- Hypoteser for én andel (p)
- Hypoteser for to andele
- Analyse af antalstabeller (χ^2-test) (Alle forventede antal > 5)

Chapter 7: Inferences for Proportions

Dansk

Statistics for proportions:

- Proportion: $p = \frac{x}{n}$ *(x successes out of n observations)*
- Specific methods: one, two and $k > 2$ samples:
  - Binary/categorical response

Specific methods:

- Estimation and confidence interval of proportions:
  - Large sample vs. small sample methods
- Hypotheses for one proportion
- Hypotheses for two proportions
- Analysis of contingency tables (χ^2-test) (All expected > 5)

Oversigt

Forskellige analyse/data-situationer

Gennemsnit for kvantitative data:

- Hypotesetest/KI for én middelværdi *(one-sample, i.e. one group/population)*
- Hypotesetest/KI for to middelværdier *(two-sample, i.e. two groups/populations)*
- Hypotesetest/KI for flere middelværdier *(k-sample, i.e. k groups/populations)*

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:

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

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

Spørgsmål om andel (`socrative.com`, ROOM: pbac)

Hvilken kan ikke en være en andel?

- A: $103/900$
- B: $12/80$
- C: 0.957
- D: $202/154$
- E: 0.224

Svar: D, x kan ikke være højere end n.

Konfidensinterval for én andel

Method 7.3

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

$$\left[\hat{p} - z_{1-\alpha/2} \cdot \hat{\sigma}_{\hat{p}}\ ,\quad \hat{p} + z_{1-\alpha/2} \cdot \hat{\sigma}_{\hat{p}}\right] \quad \left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \qquad \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

(Vi siger: Med stor sikkerhed vælger vi at tro at p i dette interval)

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, kapitel 2:

$$\mathrm{E}(X) = np$$
$$\mathrm{Var}(X) = np(1-p)$$

Derfor får man

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

Eksempel 1

Venstrehåndede:

$p =$ Andelen af venstrehåndede i Danmark

eller:

Kvindelige ingeniørstuderende:

$p =$ Andelen af kvindelige ingeniørstuderende

Eksempel 1

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

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

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

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

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

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

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

Spørgsmål om konfidensinterval fejl (`socrative.com`, ROOM: pbac)

Mulig fejl ved konfidensinterval er, at den "rigtige" værdi ikke er inkluderet i intervallet. Hvor ofte vil man begå en denne fejl ved $\alpha = 5\%$?

- A: 95% af gangene
- B: 1% af gangene
- C: 5% af gangene
- D: 50% af gangene
- E: Ved ikke

Svar: C. Der er α sandsynlighed for ikke at fange populations værdi (den "rigtige" værdi) (ligesom Type I fejl for *Hypotesetest*: H_0 er sand, men man kommer til at afvise den)

Stikprøvestørrelse: "Margin of Error" (ME):

Margin of Error på estimat kan siges at være:

- Forventningsværdi af *"halvdelen af konfidensintervallets bredde"*
- *"Den forskel i middelværdi"* man gerne vil være i stand til at påvise
- Under H_0: Forventningsværdi af afstanden mellem middelværdien og det kritiske niveau

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

Spørgsmål om Margin of Error (`socrative.com`, ROOM: pbac)

Hvad er "Margin of Error"(ME) hvis man vil have et konfidensinterval med bredde på 0.2?

- A: 0.1
- B: -0.15
- C: 0.2
- D: 0.4
- E: Ved ikke

Svar: A. ME er Forventningsværdi af *"halvdelen af konfidensintervallets bredde"*

Spørgsmål om forskel i middel (`socrative.com`, ROOM: pbac)

Hvad er "Margin of Error"(ME) hvis man vil være i stand til at påvise forskel i middelværdi på 0.2?

- A: 0.1
- B: -0.15
- C: 0.2
- D: 0.4
- E: Ved ikke

Svar: C. *"Den forskel i middelværdi"* man gerne vil være i stand til at påvise

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

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

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 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:

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

Spørgsmål om stikprøvestørrelse (socrative.com, ROOM: pbac)

Ved test af hvilken af følgende nulhypoteser skal bruges den største stikprøvestørrelse (n) ved samme α konfidens?

- A: $H_0: p = 0.2$
- B: $H_0: p = 0.1$
- C: $H_0: p = 0.4$
- D: $H_0: p = 0.95$
- E: Ved ikke

Svar: C. Jo tættere på $p = 0.5$ man kommer jo højere n

Spørgsmål om stikprøvestørrelse (socrative.com, ROOM: pbac)

Kan I nu beregne hvor mange gange man skal slå med en terning for at teste om den har sandsynlighed $1/6$ indenfor 0.01 for at slå en sekser?

- A: Ja
- B: Nej
- C: Ved ikke

Svar: Ja, det har vi lige lært, så i R:

```
## Andel (sandsynlighed) vi vil teste for
p <- 1/6
## Signifikansniveau
## (hvor ofte vil vi lave denne fejl: Terningen er fair, men
##  vi konkluderer den ikke er fair)
alpha <- 0.05
## Fejlmargen vi vil tillade
ME <- 0.01
## Beregn antal gange vi skal slå med terningen
p * (1-p) * (qnorm(1-alpha/2)/ME)^2

## [1] 5335
```


Trin ved Hypotesetest

Trin ved Hypotesetest:

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 derefter 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

Theorem 7.10 og Method 7.11

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

Test ved brug af p-værdi (Method 7.11)

Find p-værdien (bevis mod nulhypotesen):

- We only use two-sided: $2P(Z > |z_{\text{obs}}|)$ in exercises and exams
- Remark 7.9 om one-sided "less" og "greater"

Kritiske værdier

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

Eksempel 1 - fortsat

Er halvdelen af alle danskere venstrehåndede?

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

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 $\alpha = 0.05$)

Er p-værdien under 0.05? (dvs. skal nulhypotesen forkastes ved $\alpha = 0.05$)

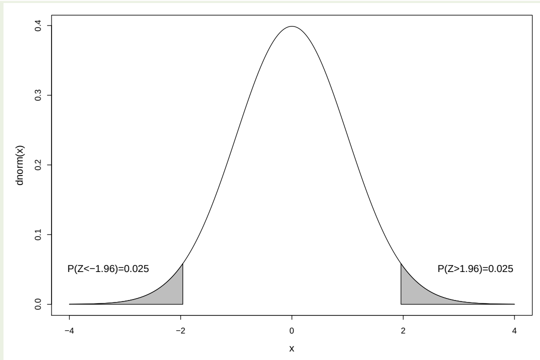
A: Ja B: Nej C: Ved ikke Svar: A

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 nulhypotesen



R: `prop.test` - een andel

```
## Single proportion

## Testing the probability = 0.5 with a two-sided alternative
## We have observed 518 out of 1154
## Without continuity corrections

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


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 $n_ip_i \geq 10$ 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:

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

Og for passende store stikprøver:

Brug standardnormalfordelingen igen

Eksempel 2

Sammenhæng mellem brug af p-piller og risikoen for blodprob i hjertet (hjerteinfarkt)

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

	p-piller	Ikke p-piller
Blodprob	23	35
Ikke blodprob	34	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 blodprob i hjertet. Anvend signifikansniveau $\alpha = 5\%$.

Eksempel 2

Sammenhæng mellem brug af p-piller og risikoen for blodprob i hjertet

	p-piller	Ikke p-piller	Sum
Blodprob	$x_1 = 23$	$x_2 = 35$	$x = 58$
Ikke blodprob	34	132	
Sum	$n_1 = 57$	$n_2 = 167$	$n = 224$

Estimater i hver stikprøve

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

R: prop.test - to andele

```
## Pill study: two proportions

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

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


Nu udfyld spørgeskema `Questionnaire` ved at tage link på hjemmesiden under `Course Material` uge12.

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 en alternativ hypotese at disse andele ikke er ens

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 (gennemsnitlig) estimat:

Under nulhypotesen fås et estimat for p

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

Hypotesetest for flere andele

Fælles (gennemsnitlig) estimat:

Under nulhypotesen fås et estimat for p

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

"Brug" dette fælles estimat i hver gruppe:

såfremt nulhypotesen gælder, vil vi forvente at den j'te gruppe har e_{1j} successer og e_{2j} fiaskoer, hvor

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

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

Hypotesetest for flere andele

Generel formel for beregning af forventede værdier i antalstabeller:

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

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 nulhypotesen

Rule of thumb for validity of the test:

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

Eksempel 2 - fortsat

De OBSERVEREDE værdier o_{ij}

	p-piller	Ikke p-piller	Total
Blodprob	23	35	
Ikke blodprob	34	132	

Eksempel 2 - fortsat

Beregn de FORVENTEDE værdier e_{ij} (altså forventede *under* H_0)

	p-piller	Ikke p-piller	Total
Blodprob			$x = 58$
Ikke blodprob			
	$n_1 = 57$	$n_2 = 167$	$n = 224$

Eksempel 2 - fortsat

Beregn de FORVENTEDE værdier e_{ij} (altså forventede *under* H_0)

	p-piller	Ikke p-piller	Total
Blodprob	14.76	43.24	$x = 58$
Ikke blodprob	42.24	123.76	
	$n_1 = 57$	$n_2 = 167$	$n = 224$

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

$$e_{12} = 167 \cdot \frac{58}{224} = 43.24$$

Eksempel 2 - fortsat

Teststørrelsen:

$$\chi^2_{\text{obs}} = \frac{(o_{11}-e_{11})^2}{e_{11}} + \frac{(o_{12}-e_{12})^2}{e_{12}} + \frac{(o_{21}-e_{21})^2}{e_{21}} + \frac{(o_{22}-e_{22})^2}{e_{22}}$$

$$=$$

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

$$= 8.33$$

Kritisk værdi og p-værdi:

```
## Kritisk værdi
qchisq(0.95, 1)

## [1] 3.8
```

```
## p-værdi
1 - pchisq(8.33, df=1)

## [1] 0.0039
```

Konklusion:

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

R: chisq.test - to andele

```
## Pill study: two proportions, chi-square test

## Chi2 test for testing the probabilities for the two groups are equal
chisq.test(pill.study, correct = FALSE)
## If we want the expected numbers save the test in an object
chi <- chisq.test(pill.study, correct = FALSE)
## The expected values
chi$expected
```


Antalstabeller

Antalstabel

- Flere end 2 kategorier (f.eks. fire.: rød, grøn, blå, sort)
- Beregningerne er ens for begge følgende setups

To mulige setups

- Setup 1: c stikprøver med r kategorier:
  - Test om der er forskel i fordelingen mellem kategorierne for hver stikprøve
- Setup 2: To kategoriske variabel (r kategorier) målt på samme individer (parret setup):
  - Test om der er forskel i fordelingen mellem de to grupper

Setup 1: c stikprøver med r kategorier

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?

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

Setup 2: To kategoriske variabel (r kategorier) målt på samme individer (parret setup)

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?

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

f.eks. er der sammenhæng mellem den måde elever klarer sig i matematik som i dansk?

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:

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

Spørgsmål (`socrative.com`, ROOM: pbac)

En 3×4 tabel - 4 stikprøver, 3-kategori udfald

	Gruppe A	Gruppe B	Gruppe C	Gruppe D	n_j
Han	3	3	2	2	10
Hun	3	3	5	2	13
Tvekøn	4	4	3	6	17
n_i	10	10	10	10	40

Hvad er e_{23}? (H_0 forventning af hunner i gruppe C)

- A: $10 \cdot 10/40$
- B: 3
- C: $10 \cdot 13/40$
- D: $17 \cdot 4/40$
- E: Ved ikke

Svar: C

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 nulhypotesen

Rule of thumb for validity of the test:

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

R: chisq.test - antalstabeller

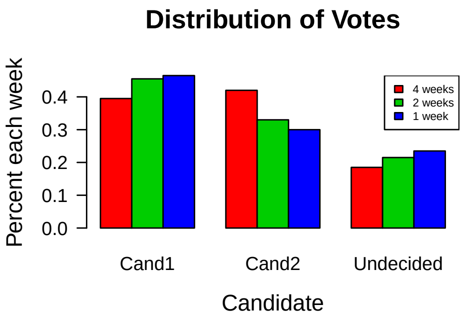
```
## Poll study: contingency table, chi-square test

## 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
```


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

```
## Testing same distribution in the three populations
chi <- chisq.test(poll, correct = FALSE)
chi
## Expected values
chi$expected
```


Oversigt