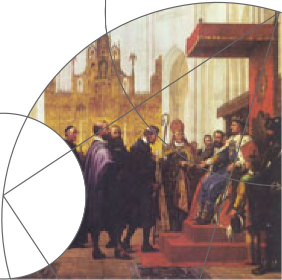
# Analyse af antalstabeller

Anders Tolver
Institut for Matematiske Fag



# I dag og næste uge

I dag: lærebogen kap. 12 (dog ikke 12.2.3, 12.2.4)

- Intro til test i tabeller
- Test for specifikke sandsynligheder
- Test for ens sandsynligheder (homogenitetstest)
- Test for uafhængighed
- Quiz 7

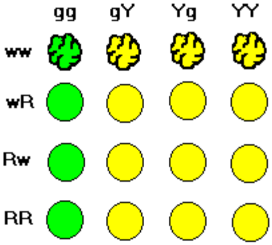
Næste uge:

- Mandag, forelæsning: repetition vha. nogle opgaver
- Mandag, øvelser: opgaveregning
- Onsdag: Ingen undervisning

# Test i tabeller

# Eksempel 12.1: Mendels ærteforsøg

| Class | Number |
|---|---:|
| Round, yellow | 315 |
| Round, green | 108 |
| Wrinkled, yellow | 101 |
| Wrinkled, green | 32 |
| Total | 556 |




- 566 ærter fra generation F2 undersøgt for farve og form
- Mendels arvelighedslære: Uafhængighed + dominans $\rightarrow$ kombinationen af fænotyper skal være forholdet $9:3:3:1$.
- Stemmer data overens med Mendels påstand?

# Eksempel: Kastrering og diabetes

Eksempel fra i mandags:

| | Diabetes | Ikke diabetes | Total |
|---|---|---|---|
| Katrerede mus | 26 | 24 | 50 |
| Ikke-kastrerede mus | 12 | 38 | 50 |

- 50 kastrerede og 50 ikke-kastrerede mus undersøgt for diabetes.
- Er sandsynlighederne for diabetes ens i to to grupper? Altså: Er proportionerne ens i de to rækker, på nær tilfældighed?
- Bemærk: Rækkesummerne kendt på forhånd (begge 50)

# Eksempel: Politik og økonomi

| | Demokrat | Republikaner | Uafhængig |
|---|---|---|---|
| Begrænse udgifter | 101 | 282 | 61 |
| Øge skatter | 38 | 67 | 25 |
| Øge offentlige invest. | 131 | 88 | 31 |
| Lade underskuddet vokse | 61 | 90 | 25 |

- 1000 tilfældige amerikanske vælgere adspurgt om to ting: politisk tilhørsforhold og foretrukne finanspolitisk instument
- Er de to ting uafhængige?
- Bemærk: De 1000 personer er udtrukket tilfældigt. Hverken række- eller søjlesummer, kun totalsummen, kendt på forhånd.

# Ligheder og forskelle mellem dataeksemplerne

Data:

- I alle tre eksempler kunne vi beskrive data vha. en **antalstabel** (eng.: contingency table)
- Interesseret i specifikke cellesandsynligheder (Mendel) eller sammenhænge mellem cellesandsynligheder (de andre eks.)
- I tovejstabellerne: Rækkesummer kendte (diabetes) eller kun totalsummen kendt (politik)

# Ligheder og forskelle mellem dataeksemplerne

Data:

- I alle tre eksempler kunne vi beskrive data vha. en **antalstabel** (eng.: contingency table)
- Interesseret i specifikke cellesandsynligheder (Mendel) eller sammenhænge mellem cellesandsynligheder (de andre eks.)
- I tovejstabellerne: Rækkesummer kendte (diabetes) eller kun totalsummen kendt (politik)

Hypotese afhænger af dataindsamlingen:

- Test for **specifikke sandsynligheder** (goodness-of-fit)
- Test for **ens sandsynligheder/proportioner** (homogenitetstest)
- Test for **uafhængighed**

# Hypotesetest i antalstabeller

I alle tilfælde:

- Beregn **forventet antal obs.** i hver celle under hypotesen
- Beregn **teststørrelse**

$$X^2_{\text{obs}} = \sum_{\text{alle celler}} \frac{(\text{observeret} - \text{forventet})^2}{\text{forventet}}$$

  $X^2_{\text{obs}}$ måler forskellen mellem tabel med observerede værdier og tabel med forventede værdier.

- Bestem ***p*-værdi** ved at sammenligne $X^2_{\text{obs}}$ med en (den rigtige) $\chi^2$-fordeling. Detaljer kommer senere.

Er tabellerne med obs. hhv. forventede antal så forskellige at det må skyldes at hypotesen er falsk, eller kan det skyldes tilfældigheder?

# Goodness-of-fit test (GOF): Test for specifikke sandsynligheder

# Mendels ærteforsøg: Model og hypotese

| Class | Number |
|---|---:|
| Round, yellow | 315 |
| Round, green | 108 |
| Wrinkled, yellow | 101 |
| Wrinkled, green | 32 |
| Total | 556 |

**Stat. model:** $n = 556$ uafhængige obs. der hver især kan havne i $k = 4$ grupper; alle med (ukendte) sandsynligheder $p_1, \ldots, p_k$.

# Mendels ærteforsøg: Model og hypotese

| Class | Number |
|---|---|
| Round, yellow | 315 |
| Round, green | 108 |
| Wrinkled, yellow | 101 |
| Wrinkled, green | 32 |
| Total | 556 |

**Stat. model:** $n = 556$ uafhængige obs. der hver især kan havne i $k = 4$ grupper; alle med (ukendte) sandsynligheder $p_1, \ldots, p_k$.

**Hypotese**,

$$H_0 : p_1 = \frac{9}{16}, \quad p_2 = \frac{3}{16}, \quad p_3 = \frac{3}{16}, \quad p_4 = \frac{1}{16}$$

Generelt: $H_0 : p_1 = p_{01}, \ldots, p_k = p_{0,k}$ for **kendte ssh,** $p_{01}, \ldots, p_{0k}$.

# Mendels ærteforsøg: Forventede værdier

**Hvis** hypotesen er sand, hvor mange observationer ville vi så **forvente** i hver gruppe?

$$E_i = \text{expected}_i = n \cdot p_{i0}$$

# Mendels ærteforsøg: Forventede værdier

**Hvis** hypotesen er sand, hvor mange observationer ville vi så **forvente** i hver gruppe?

$$E_i = \text{expected}_i = n \cdot p_{i0}$$

For Mendels data:

| Class | Observed | Expected |
|---|---:|---:|
| Round, yellow | 315 | 312.75 |
| Round, green | 108 | 104.25 |
| Wrinkled, yellow | 101 | 104.25 |
| Wrinkled, green | 32 | 34.75 |
| Total | 556 | 556 |

# Mendels ærteforsøg: Teststørrelse og $p$-værdi

Teststørrelse:

$$\begin{aligned} X^2_{\mathrm{obs}} &= \sum_{i=1}^{4} \frac{(\mathsf{observed}_i - \mathsf{expected}_i)^2}{\mathsf{expected}_i} \\ &= \frac{(315-312.75)^2}{312.75} + \cdots + \frac{(32-34.75)^2}{34.75} = 0.470 \end{aligned}$$

$X^2$ er altid $\geq 0$, og **store værdier passer dårligt med $H_0$** (er kritiske), små værdier passer godt med $H_0$.

# Mendels ærteforsøg: Teststørrelse og *p*-værdi

Teststørrelse:

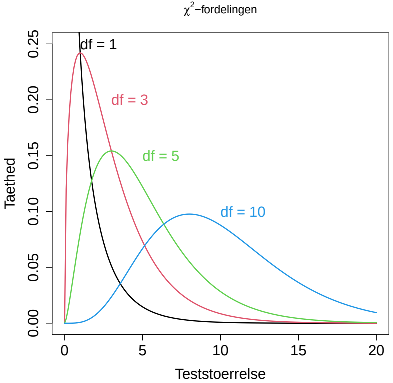
$$X^2_{\text{obs}} = \sum_{i=1}^{4} \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i}$$
$$= \frac{(315 - 312.75)^2}{312.75} + \cdots + \frac{(32 - 34.75)^2}{34.75} = 0.470$$

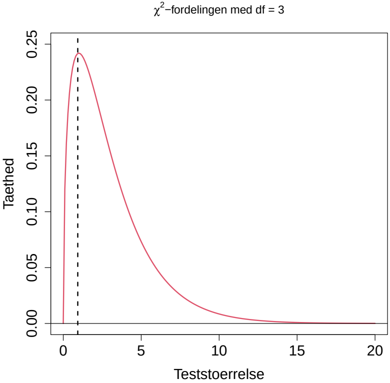
$X^2$ er altid $\geq 0$, og **store værdier passer dårligt med $H_0$** (er kritiske), små værdier passer godt med $H_0$.

*p*-værdi:

- Sandsynlighed for at få værdi af $X^2$ der er $\geq X^2_{\text{obs}}$
- Viser sig at *p*-værdien skal bestemmes i **$\chi^2$-fordelingen** (chi-i-anden) med $k - 1 = 4 - 1 = 3$ frihedsgrader.

# $\chi^2$-fordelinger, beregning af $p$-værdi




- $p$-værdien er arealet **til højre for** $X^2_{\mathrm{obs}}$
- Her fås $p$-værdien 0.93, så vi accepterer hypotesen

# R: chisq.test

```
### Testet
chisq.test(c(315,108,101,32), p=c(9,3,3,1)/16)

##
##  Chi-squared test for given probabilities
##
## data:  c(315, 108, 101, 32)
## X-squared = 0.47002, df = 3, p-value = 0.9254

### De forventede vaerdier
chisq.test(c(315,108,101,32), p=c(9,3,3,1)/16)$expected

## [1] 312.75 104.25 104.25  34.75
```

# Mendels ærteforsøg: Opsummering

- Stat. model: 556 uafhængige obs. der hver især kan havne i 4 grupper; alle med (ukendte) sandsynligheder $p_1, \dots, p_4$.
- Hypotese, svarende til Mendels love:

$$p_1 = \frac{9}{16},\ p_2 = \frac{3}{16},\ p_3 = \frac{3}{16},\ p_4 = \frac{1}{16}$$

- $\chi^2$-test gav $p = 0.93$ ($X^2_{\mathrm{obs}} = 0.47$, $\mathrm{df} = 3$)
- Hypotesen accepteres, så data er i fin overensstemmelse med Mendels teorier

# Test for ens sandsynligheder/proportioner: Homogenitetstest

# Eksempel: Kastrering og diabetes

| | Diabetes | Ikke diabetes | Total |
|---|---|---|---|
| Katrerede mus | 26 | 24 | 50 |
| Ikke-kastrerede mus | 12 | 38 | 50 |

- **Rækkesummer kendt på forhånd**. Kunne have organiseret data det i stedet var søjlesummerne der var kendt på forhånd.
- I hver række har vi sandsynligheder for at havne i hver søjle. **For hver række summerer sandsynlighederne til 1.**
- Vi er interesseret i **om sandsynligheden for diabetes er ens for kastrerede og ikke-kastrerede mus**
- Der kunne være flere rækker og/eller søjler

# Homogenitetstest: Generel notation

| | søjle 1 | søjle 2 | $\cdots$ | søjle $k$ | Total |
|---|---|---|---|---|---|
| række 1 | $y_{11}$ | $y_{12}$ | $\cdots$ | $y_{1k}$ | $n_1$ |
| række 2 | $y_{21}$ | $y_{22}$ | $\cdots$ | $y_{2k}$ | $n_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| række $r$ | $y_{r1}$ | $y_{r2}$ | $\cdots$ | $y_{rk}$ | $n_r$ |
| Total | $s_1$ | $s_2$ | $\cdots$ | $s_k$ | $n$ |

Dataindsamling:

- $r$ **populationer** (rækker), $n_i$ observationer fra population $i$
- I hver population er observationerne klassificeret efter et kriterium med $k$ muligheder.
- Rækkesummer (men ikke søjlesummer) kendt på forhånd.

# Homogenitetestest: Sandsynligheder og hypotese

| | søjle 1 | søjle 2 | $\cdots$ | søjle $k$ | Total |
|---|---|---|---|---|---|
| række 1 | $p_{11}$ | $p_{12}$ | $\cdots$ | $p_{1k}$ | 1 |
| række 2 | $p_{21}$ | $p_{22}$ | $\cdots$ | $p_{2k}$ | 1 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| række $r$ | $p_{r1}$ | $p_{r2}$ | $\cdots$ | $p_{rk}$ | 1 |

Hypotesen er at sandsynlighederne/proportionerne er ens i alle populationer:

$$p_{1j} = p_{2j} = \cdots = p_{rj} \text{ for alle søjler } j$$

Altså at fordelingen henover søjlerne er den samme for alle rækker.

Hvis der kun er to søjler: Sammenligning af $r$ binomialfordelinger

# Homogenitetstest: Statistisk model og hypotese

**Statistisk model:**

- Uafhængige obs. fra $r$ populationer med $n_i$ obs. i population $i$. Hver obs. kan havde i $k$ grupper/celler
- I population $i$ er sandsynligheden for at havne i gruppe $j$ lig $p_{ij}$. Summen af $p_{ij}$'erne er 1 for hvert $i$ for sig
- Hvis der kun er to søjler: $r$ binomialfordelinger

# Homogenitetstest: Statistisk model og hypotese

**Statistisk model:**

- Uafhængige obs. fra $r$ populationer med $n_i$ obs. i population $i$. Hver obs. kan havde i $k$ grupper/celler
- I population $i$ er sandsynligheden for at havne i gruppe $j$ lig $p_{ij}$. Summen af $p_{ij}$'erne er 1 for hvert $i$ for sig
- Hvis der kun er to søjler: $r$ binomialfordelinger

**Hypotesen om homogenitet** er at søjlesandsynlighederne er ens for alle rækker:

$$p_{1j} = p_{2j} = \cdots = p_{rj} \text{ for alle søjler } j.$$

**To søjler: Sammenligning af binomialsandsynligheder!**

# Homogenitetstest: Forventede værdier

Under hypotesen **estimeres søjlesandsynlighederne** — fælles for alle rækker — naturligt som

$$\hat{q}_j = \frac{s_j}{n} = \frac{\text{søjlesum}_j}{n}$$

# Homogenitetstest: Forventede værdier

Under hypotesen **estimeres søjlesandsynlighederne** — fælles for alle rækker — naturligt som

$$\hat{q}_j = \frac{s_j}{n} = \frac{\text{søjlesum}_j}{n}$$

**Forventet antal** i celle $(i, j)$ hvis hypotesen er sand:

$$E_{ij} = n_i \cdot \hat{q}_j = \frac{\text{rækkesum}_i \cdot \text{søjlesum}_j}{\text{totalsum}}$$

# Kastrering og diabetes: Forventede værdier

Data:

| | Diabetes | Ikke diabetes | Total |
|---|---|---|---|
| Katrerede mus | 26 | 24 | 50 |
| Ikke-kastrerede mus | 12 | 38 | 50 |

Forventede værdier:

| | Diabetes | Ikke diabetes | Total |
|---|---|---|---|
| Katrerede mus | 19 | 31 | 50 |
| Ikke-kastrerede mus | 19 | 31 | 50 |

# Kastrering og diabetes: Teststørrelse og $p$-værdi

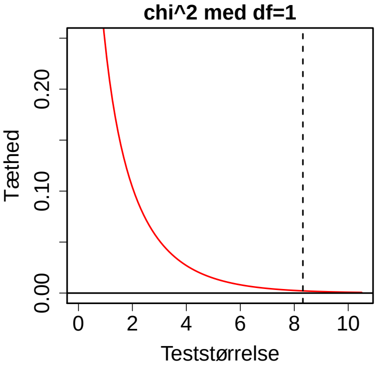
Teststørrelse:

$$
\begin{aligned}
X^2_{\text{obs}} &= \sum_{\text{alle celler}} \frac{(y_{ij} - E_{ij})^2}{E_{ij}} \\
&= \sum_{\text{alle celler}} \frac{(\text{observed}_{ij} - \text{expected}_{ij})^2}{\text{expected}_{ij}} \\
&= \frac{(26-19)^2}{19} + \frac{(24-31)^2}{31} + \frac{(12-19)^2}{19} + \frac{(38-31)^2}{31} \\
&= 8.32
\end{aligned}
$$

Store værdier passer dårligt med $H_0$, små værdier passer godt.

$p$-værdi: Viser sig at $X^2_{\text{obs}}$ skal vurderes i $\chi^2$-fordelingen med $\text{df} = (r-1)(k-1) = 1$

# Kastrering og diabetes: $\chi^2$-fordelingen og *p*-værdien




- *p*-værdien er arealet **til højre for** den $X^2_{\mathrm{obs}}$
- Her fås *p*-værdien 0.0039, så hypotesen forkastes klart

# R: chisq.test

```
diabetes <- matrix(c(26,12,24,38), 2,2)
diabetes

##      [,1] [,2]
## [1,]   26   24
## [2,]   12   38

chisq.test(diabetes, correct=FALSE)

##
##  Pearson's Chi-squared test
##
## data:  diabetes
## X-squared = 8.3192, df = 1, p-value = 0.003923

chisq.test(diabetes, correct=FALSE)$expected

##      [,1] [,2]
## [1,]   19   31
## [2,]   19   31
```

# R: prop.test

```
prop.test(c(26,12), c(50,50), correct=FALSE)

##
##  2-sample test for equality of proportions without continuity
##
## data:  c(26, 12) out of c(50, 50)
## X-squared = 8.3192, df = 1, p-value = 0.003923
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.09781821 0.46218179
## sample estimates:
## prop 1 prop 2
##   0.52   0.24
```

# Kastrering og diabetes: Opsummering

- Statistisk model: Data fra to binomialfordelinger med successandsynligheder $p_{11}$ og $p_{21}$
- Hypotese om homogenitet, $H_0 : p_{11} = p_{21}$. Vi fik $p = 0.0039$, så hypotesen afvises.

  Der er forskel på risikoen for at udvikle diabetes.
- Kastrering øger risikoen for diabetes: Forskellen mellem ssh. estimeres til 0.280 med 95% KI $(0.098\,,\,0.462)$

# Test for uafhængighed

# Studerende på StatData1 i 2022

Ved forelæsningen i StatData1 d. 7/9-2022 svarede 130 studerede bl.a. på følgende spørgsmål

- Glæder du dig til kurset Statistisk Dataanalyse 1? (Ja/Nej)
- Giv et realitisk bud på din karakter ved eksamen i Statistisk Dataanalyse 1 (her grupperet: 02-4, 7, 10-12)

```
##
##       02-4 10-12  7
##   Ja    25    31 33
##   Nej   19     4 18
```

- Hverken række- eller søjlesummer kendt på forhånd.
- Er svarene på de to spørgsmål uafhængige? ... Hvad skal det egentlig betyde?

# Forventing til SD1 og til eksamenskarakter

**Hypotese:** Ingen sammenhæng mellem forventning til SD1 og forventning til eget eksamensresultat.

For eksempel:

$$P(\text{02-4 } \textbf{og} \text{ Ja}) = P(\text{02-4}) \cdot P(\text{Ja})$$

Altså at sandsynligheden for at begge dele er opfyldt fås ved at **gange** de to sandsynligheder. Skal gælde for **alle celler** i tabellen.

Hvis $p_{ij}$ er cellesandsynligheder, $p_i$ er rækkesandsynligheder og $q_j$ er søjlesandsynligheder er hypotesen at

$$p_{ij} = p_i \cdot q_j \text{ for alle } i, j$$

# Uafhængighedstest: Generel notation

| | søjle 1 | søjle 2 | $\cdots$ | søjle $k$ | Total |
|---|---|---|---|---|---|
| række 1 | $y_{11}$ | $y_{12}$ | $\cdots$ | $y_{1k}$ | $n_1$ |
| række 2 | $y_{21}$ | $y_{22}$ | $\cdots$ | $y_{2k}$ | $n_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| række $r$ | $y_{r1}$ | $y_{r2}$ | $\cdots$ | $y_{rk}$ | $n_r$ |
| Total | $s_1$ | $s_2$ | $\cdots$ | $s_k$ | $n$ |

- Alle observationer **klassificeret efter to kriterier**. Organiseret i $r$ rækker og $k$ søjler
- **Kun totalsumen $n$ er kendt på forhånd**
- Rækkesummer og søjlesummer ikke kendt på forhånd, men kan selvfølgelig beregnes når vi har data

# Uafhængighedstest: Statistisk model

**Statistisk model:**

- $n$ uafhængige obs. der hver især kan havne i $r \cdot k$ celler
- Ssh. for celle $(i, j)$ kaldes $p_{ij}$. Sum af **alle** $p_{ij}$'er er 1

Rækkesandsynligheder $p_i$ og søjlesandsynligheder $q_j$. Sum af de relevante cellesandsynligheder.

| | søjle 1 | søjle 2 | $\cdots$ | søjle $k$ | Total |
|---|---|---|---|---|---|
| række 1 | $p_{11}$ | $p_{12}$ | $\cdots$ | $p_{1k}$ | $p_1$ |
| række 2 | $p_{21}$ | $p_{22}$ | $\cdots$ | $p_{2k}$ | $p_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| række $r$ | $p_{r1}$ | $p_{r2}$ | $\cdots$ | $p_{rk}$ | $p_r$ |
| Total | $q_1$ | $q_2$ | $\cdots$ | $q_k$ | 1 |

# Uafhængighedstest: Hypotese

**Hypotese om uafhængighed:**

$$\begin{aligned} p_{ij} &= \text{Sandsynlighed for række } i \text{ og søjle } j \\ &= \text{Sandsynlighed for række } i \cdot \text{Sandsynlighed for søjle } j \\ &= p_i \cdot q_j \end{aligned}$$

Hypotesen er at dette gælder for **alle** $i$ og $j$, dvs. alle celler.

# Forventede værdier

Estimater for række- og søjlesandsynligheder:

$$\hat{p}_i = \frac{\text{rækkesum}_i}{\text{totalsum}} = \frac{n_i}{n}, \quad \hat{q}_j = \frac{\text{søjlesum}_j}{\text{totalsum}} = \frac{s_j}{n}$$

Under hypotesen har vi derfor følgende **estimater for cellessh.:**

$$\hat{p}_{ij} = \hat{p}_i \cdot \hat{q}_j = \frac{\text{rækkesum}_i \cdot \text{søjlesum}_j}{n^2}$$

# Forventede værdier

Estimater for række- og søjlesandsynligheder:

$$\hat{p}_i = \frac{\text{rækkesum}_i}{\text{totalsum}} = \frac{n_i}{n}, \quad \hat{q}_j = \frac{\text{søjlesum}_j}{\text{totalsum}} = \frac{s_j}{n}$$

Under hypotesen har vi derfor følgende **estimater for cellessh.:**

$$\hat{p}_{ij} = \hat{p}_i \cdot \hat{q}_j = \frac{\text{rækkesum}_i \cdot \text{søjlesum}_j}{n^2}$$

**Forventet antal** i celle $(i, j)$ hvis $H_0$ er sand:

$$E_{ij} = n \cdot \hat{p}_{ij} = \frac{\text{rækkesum}_i \cdot \text{søjlesum}_j}{\text{totalsum}}$$

Præcis det **samme som for homogenitetstestet!**

# SD1 og eksamenskarakter: Forventede værdier

Data:

| | 02-4 | 7 | 10-12 | I alt |
|---|---|---|---|---|
| Ja | 25 | 33 | 31 | 89 |
| Nej | 19 | 18 | 4 | 41 |
| I alt | 44 | 51 | 35 | 130 |

Forventede værdier:

| | 02-4 | 7 | 10-12 | I alt |
|---|---|---|---|---|
| Ja | 30.12 | 34.92 | 23.96 | 89 |
| Nej | 13.88 | 16.08 | 11.04 | 41 |
| I alt | 44 | 51 | 34 | 130 |

# SD1 og eksamenskarakter: Teststørrelse og *p*-værdi

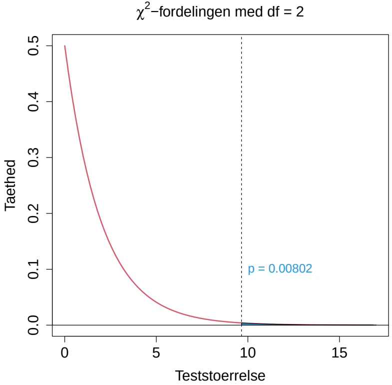
Teststørrelse

$$
\begin{aligned}
X^2_{\mathrm{obs}} &= \sum_{\text{alle celler}} \frac{(y_{ij} - E_{ij})^2}{E_{ij}} \\
&= \sum_{\text{alle celler}} \frac{(\text{observed}_{ij} - \text{expected}_{ij})^2}{\text{expected}_{ij}} \\
&= \frac{(25 - 30.12)^2}{30.12} + \cdots + \frac{(4 - 11.04)^2}{11.04} \\
&= 9.651
\end{aligned}
$$

Store værdier passer dårligt med $H_0$, små værdier passer godt.

*p*-værdi: Viser sig at $X^2_{\mathrm{obs}}$ skal vurderes i $\chi^2$-fordelingen med $\mathrm{df} = (r-1)(k-1) = (2-1)\cdot(3-1) = 2$. **Ligesom homogenitetstestet!**

# SD1 og eksamenskarakter: $\chi^2$-fordelingen og $p$-værdien




- $p$-værdien er arealet **til højre for** $X^2_{\text{obs}}$
- Her fås stort set en $p$-værdi på 0, så hypotesen forkastes

# R: chisq.test

```
sd1data <- matrix(c(25, 19, 33, 18, 31, 4), 2, 3)
sd1data

##      [,1] [,2] [,3]
## [1,]   25   33   31
## [2,]   19   18    4

chisq.test(sd1data, correct = FALSE)

##
##  Pearson's Chi-squared test
##
## data:  sd1data
## X-squared = 9.6512, df = 2, p-value = 0.008022
```

# R: chisq.test

```
### De forventede vaerdier
chisq.test(sd1data)$expected

##             [,1]     [,2]     [,3]
## [1,] 30.12308 34.91538 23.96154
## [2,] 13.87692 16.08462 11.03846
```

# SD1 og eksamenskarakter: Opsummering

- Stat. model: 130 uafhængige obs. der hver især kan havne i 6 grupper; alle med (ukendte) sandsynligheder $p_{ij}$
- Hypotese om uafhængighed: $p_{ij} = p_i \cdot q_j$ for alle $i, j$
- $\chi^2$-test gav $p = 0.008022$ ($X^2_{\mathrm{obs}} = 9.6512$, $\mathrm{df} = 2$)
- Hypotesen forkastes, så forventningen til kurset StatData1 og til eksamensresultatet er IKKE uafhængige

# Diverse

# Uafhængighedstest vs. homogenitetstest

**Beregningerne er helt identiske:**

- Forventede værdier beregnes som $E_{ij} = \frac{\text{rækkesum}_i \cdot \text{ søjlesum}_j}{\text{totalsum}}$
- Teststørrelse beregnes som $X^2_{\mathrm{obs}} = \sum \frac{(\text{observeret}-\text{forventet})^2}{\text{forventet}}$
- Teststørrelsen vurderes i $\chi^2$-ford. med $\mathrm{df} = (r-1)(k-1)$: $p$-værdien beregnes som sandsynlighed til højre for $X^2_{\mathrm{obs}}$
- Hypotesen forkastes/afvises på baggrund af $p$-værdien som sædvanlig
- Testet kan udføres med `chisq.test` i R

# Uafhængighedstest vs. homogenitetstest

**Beregningerne er helt identiske:**

- Forventede værdier beregnes som $E_{ij} = \frac{\text{rækkesum}_i \cdot \text{ søjlesum}_j}{\text{totalsum}}$
- Teststørrelse beregnes som $X^2_{\mathrm{obs}} = \sum \frac{(\text{observeret} - \text{forventet})^2}{\text{forventet}}$
- Teststørrelsen vurderes i $\chi^2$-ford. med $\mathrm{df} = (r-1)(k-1)$: $p$-værdien beregnes som sandsynlighed til højre for $X^2_{\mathrm{obs}}$
- Hypotesen forkastes/afvises på baggrund af $p$-værdien som sædvanlig
- Testet kan udføres med `chisq.test` i R

Men: **Hypotesen og derfor fortolkningen er forskellig** afhængig af datastrukturen/-indsamlingen.

# Uafhængighedstest vs. homogenitetstest

Uafhængighedstest:

- Når **to kategoriske variable** med hhv. $r$ og $k$ kategorier er observeret for **en enkelt population**
- Hverken række- eller søjlesummer er kendt på forhånd
- Hypotese om **uafhængighed** mellem de to variable

# Uafhængighedstest vs. homogenitetstest

Uafhængighedstest:

- Når **to kategoriske variable** med hhv. $r$ og $k$ kategorier er observeret for **en enkelt population**
- Hverken række- eller søjlesummer er kendt på forhånd
- Hypotese om **uafhængighed** mellem de to variable

Homogenitetstest:

- Når **en enkelt kategorisk variabel** med $k$ kategorier er observeret i $r$ **forskellige populationer**
- Rækkesummer (eller søjlesummer) kendt på forhånd
- Hypotese om **ens proportioner/sandsynligheder** for de $r$ populationer

# Kontinuitetskorrektion

For $2 \times 2$ tabeller (men ikke større tabeller) laver `chisq.test` som default en **kontinuitetskorrektion**, når $X^2$ beregnes.

- `chisq.test(..., correct=FALSE)`: Giver det vi netop har beregnet
- `chisq.test(..., correct=TRUE)`: Giver lidt andre resultater — faktisk forbedret.

Begge dele er OK til eksamen, medmindre der står noget specifikt.

# R: Med og uden kontinuitetskorrektion

```
diabetes

##      [,1] [,2]
## [1,]   26   24
## [2,]   12   38

chisq.test(diabetes, correct=TRUE)

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  diabetes
## X-squared = 7.1732, df = 1, p-value = 0.0074

chisq.test(diabetes, correct=FALSE)

##
##  Pearson's Chi-squared test
##
## data:  diabetes
## X-squared = 8.3192, df = 1, p-value = 0.003923
```

# SD1 og eksamenskarakter: R warning

Man kan komme ud for, at `chisq.test` giver en advarsel!

Her benyttes alle trin på karakterskalaen

```
new_sd1data

##
##       10 12  2  4  7
##   Ja  26  5  7 18 33
##   Nej  4  0  5 14 18

chisq.test(new_sd1data, correct = FALSE)

## Warning in chisq.test(new_sd1data, correct = FALSE): Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  new_sd1data
## X-squared = 10.022, df = 4, p-value = 0.04007
```

# Approksimation

Vi har hele tiden sagt at $X^2$ kommer fra $\chi^2$-fordeling når hypotesen er sand, men faktisk er det kun en approksimation.

Tommelfingerregel: **Approksimationen er kun god hvis de forventede værdier i alle celler er $\geq 5$.**

```
chisq.test(new_sd1data)$expected

## Warning in chisq.test(new_sd1data):  Chi-squared
approximation may be incorrect

##
##              10       12        2        4        7
##   Ja  20.538462 3.423077 8.215385 21.90769 34.91538
##   Nej  9.461538 1.576923 3.784615 10.09231 16.08462
```

# Hvad gør man hvis forventede antal er for små?

- **Slå rækker og/eller søjler** sammen så tommelfingerreglen om forventede værdier er OK.

  Sammenlægningen skal selvfølgelig give mening, typisk for ordinale data. (Kunne godt gøres her!)

- Beregn *p*-værdien ved **simulation.**

  Laver mange datasæt som de ville se ud hvis hypotesen var sand og beregner $X^2$. Hvor ofte er den større end $X^2_{\mathrm{obs}}$?

# R: Simuleret *p*-værdi

```
set.seed(2022)
chisq.test(new_sd1data, simulate.p.value = TRUE, B=10000)

##
##  Pearson's Chi-squared test with simulated p-value (based on 1
##  replicates)
##
## data:  new_sd1data
## X-squared = 10.022, df = NA, p-value = 0.0341
```

# SD1 og eksamenskarakter: Konklusion

- Simulerede $p$-værdier er lidt forskellige fra gang til gang (medmindre man som her vælger fast `seed`)
- De simulerede $p$-værdier er tæt på $p$-værdien baseret på $\chi^2$-approksimationen (0.04007)
- Tegn på sammenhæng ml. forventning til SD1 og forventning til eksamenskarakter

# Opsummering vedr. R

Test i tabeller:

- `chisq.test`: Giver $X^2_{\text{obs}}$ og $p$-værdi samt forventede værdier. Kan også beregne simulerede $p$-værdier. Ingen konfidensint.
- `prop.test`: Kan bruges hvis der kun er to søjler (evt. flere rækker). Giver ikke de forventede værdier.

  Også KI for forskel mellem rækkessh. for $2 \times 2$ tabeller.
- For $2 \times 2$ tabeller: `chisq.test` og `prop.test` fås med/uden kontinuitetskorrektion.
- Data skal indtastes forskelligt når man bruger `chisq.test` og `prop.test`.

Vælg selv metoden medmindre du bliver spurgt om noget eksplicit.