# HD Dataanalyse

## *Note 8: Analyse af mere end to grupper*

Copenhagen Business School

# Emne i dette notesæt

Når vi betragter ***to variable*** i et datamateriale, er vi interesseret i at vurdere, om der er en ***sammenhæng mellem dem*** eller ej. Vurderingen foretager vi ved hjælp af et passende valgt hypotesetest.

Hvis den *ene* af de to variable er kategorisk, tænker vi ofte på variablens værdier som "grupper", der inddelerer/grupperer værdierne af den *anden* variabel.

Vi har tidligere set på tilfældet med to grupper (i note 7b). Vi skal i denne note se på det mere generelle tilfælde, hvor den ene variabel har ***et endeligt antal*** ($\geq 2$) mulige værdier/grupper.

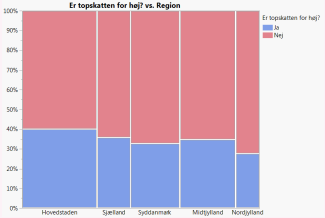
Helt overordnet, så beskriver noterne 7b, 8 og 9 hver især en metode til hypotesetest af en sammenhæng mellem to variable. Hvilken metode/note vi skal bruge afhænger af, hvilke mulige værdier de to variable har.

| Variabel 1 \ Variabel 2 | *2 mulige værdier* | *Endeligt antal mulige værdier* | *Mange mulige værdier* |
|---|---|---|---|
| *2 mulige værdier* | Test for to andele $p_1$ og $p_2$ (Note 7b) | Test for uafhængighed (Note 8) | Test for to middelværdier $\mu_1$ og $\mu_2$ (Note 7b) |
| *Endeligt antal mulige værdier* | | | Test for et endeligt antal middelværdier (Note 8) |
| *Mange mulige værdier* | | | Regressionsanalyse med én forklarende var. (Note 9) |

# Emne i dette notesæt

**Eksempel: Skat**

I spørgeskemaundersøgelsen om velfærd og skat ser vi på spørgsmålet *"Er topskatten for høj?"*. Figuren nedenfor viser de afgivne svar inddelt efter hvilken af landets fem regioner (= "grupper"), de adspurgte personer er bosidende i. Det ser ud til, at der med undtagelse af personer bosiddende i Nordjylland, ikke er væsentlig forskel på opfattelsen af, om topskatten er for høj:




Spørgsmålet er, om der på baggrund af datamaterialet er belæg for en påstand om, at der er forskel på andelen af befolkningen, som mener topskatten er for høj, i de enkelte regioner? Eller om de observerede forskelle ikke er større end, hvad der kan forklares med tilfældig variation blandt de adspurgte personer?

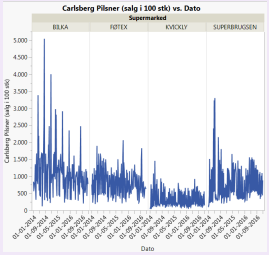
Med andre ord, skal vi forvente samme andel af befolkningen, der mener topskatten er for høj, i de fem regioner eller skal vi ikke? (dvs. er der en sammenhæng mellem de to variable *"Er topskatten for høj?"* og *"Region"* eller er der ikke?)

▶ JMP-video [Graph -> Graph Builder]

# Emne i dette notesæt

**Eksempel: Ølsalg**

Figuren nedenfor viser den ugentlige omsætning af Carlsberg (1 stk. 33 cl glasflaske) i de fire supermarkedskæder (= "grupper") Bilka, Føtex, Kvickly og SuperBrugsen. Det ser ud til, at der sælges nogenlunde lige mange Carlsberg i Bilka, Føtex og SuperBrugsen, mens der sælges lidt færre i Kvickly:




Spørgsmålet er, om der på baggrund af datamaterialet er belæg for en påstand om, at det forventede ugentlige salg af Carlsberg er forskelligt i de fire supermarkedskæder? Eller om de observerede forskelle i datamaterialet ikke er større end, hvad der kan forklares med tilfældig variation i omsætningen fra uge til uge?

Med andre ord, skal vi forvente samme ugentlige omsætning af Carlsberg i de fire kæder eller skal vi ikke? (dvs. er der en sammenhæng mellem variablene *"Carlsberg Pilsner (salg i 100 stk)"* og *"Supermarked"* eller er der ikke?)

JMP-video [Graph -> Graph Builder]

# Indholdsfortegnelse

# 1 Hypotesetest om uafhængighed

*Hvad er uafhængighed?* • *Hypotesetest* • *$\chi^2$-fordelingen*

# 2 Hypotesetest om et antal middelværdier

# 3 OPSUMMERING

## HVAD ER UAFHÆNGIGHED?



Når vi i denne note taler om "hypotesetest om uafhængighed" betragter vi et datamateriale med observationer af ***to kategoriske variable*** med henholdsvis $R$ ("row" = række) og $C$ ("column" = søjle) mulige værdier.

Fordi begge variable er kategoriske kan vi tænke på det som om...

- datamaterialet er grupperet efter værdierne i de $R$ rækker
- datamaterialet er grupperet efter værdierne i de $C$ søjler

og det er principielt ligegyldigt, om vi gør det ene eller det andet.

Tilfældet $R = C = 2$ svarer til analyse af to andele $p_1$ og $p_2$ i note 7b:

- Den ene variabel angiver om en observation hører til gruppe 1 eller gruppe 2
- Den anden variabel angiver om en observation har værdien "1" eller "0" (de to mulige værdier for en variabel, når vi ser på en andel $p$)

## HVAD ER UAFHÆNGIGHED?



For overskuelighedens skyld opsummerer vi datamaterialet i en krydstabel med den ene variabels $R$ mulige værdier i rækkerne og den anden variabels $C$ mulige værdier i søjlerne.

Hvis vi lader $X_{r,c}$ betegne værdien i tabellens $r$'te række og $c$'te søjle, får krydstabellen følgende udseende:

| | | *Variabel 2* | | |
|---|---|---|---|---|
| *Variabel 1* | $X_{1,1}$ | $X_{1,2}$ | $\dots$ | $X_{1,C}$ |
| | $X_{2,1}$ | $X_{2,2}$ | $\dots$ | $X_{2,C}$ |
| | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| | $X_{R,1}$ | $X_{R,2}$ | $\dots$ | $X_{R,C}$ |

# Hvad er uafhængighed?



**Eksempel: Skat**

I spørgeskemaundersøgelsen om velfærd og skat har vi de to kategoriske variable *Region* og *"Er topskatten for høj?"*.

Laver vi en krydstabel mellem de to variable, og placerer variablen *Region* i rækkerne og variablen *"Er topskatten for høj?"* i søjlerne får tabellen nedenstående udseende.

**Contingency Table**

| Region | Count / Expected | Ja | Nej | Total |
|---|---|---|---|---|
| | | Er topskatten for høj? | | |
| Hovedstaden | Count | 124 | 185 | 309 |
| | Expected | 109,655 | 199,345 | |
| Sjælland | Count | 49 | 88 | 137 |
| | Expected | 48,6174 | 88,3826 | |
| Syddanmark | Count | 66 | 136 | 202 |
| | Expected | 71,6841 | 130,316 | |
| Midtjylland | Count | 80 | 149 | 229 |
| | Expected | 81,2656 | 147,734 | |
| Nordjylland | Count | 27 | 71 | 98 |
| | Expected | 34,7774 | 63,2226 | |
| Total | | 346 | 629 | 975 |

Variablen *Region* har $R = 5$ mulige værdier, mens variablen *"Er topskatten for høj?"* har $C = 2$ mulige værdier.

Af tabellen kan man eksempelvis se, at af de i alt 975 adspurgte personer i undersøgelsen, er de 309 bosiddende i Region Hovedstaden og blandt disse mener 124, at topskatten er for høj.

JMP-video [Analyze -> Fit Y by X]

## HVAD ER UAFHÆNGIGHED?



Vi er interesseret i at undersøge, om der er en sammenhæng mellem krydstabellens to variable, dvs. undersøge om der er en sammenhæng mellem variablens rækker og søjler.

HVIS der IKKE er en sammenhæng betyder det, at tabellens to variable opfører sig ***uafhængigt*** af hinanden. Vi benytter derfor begreberne "sammenhæng/ingen sammenhæng" henholdsvis "afhængighed/uafhængighed" synonymt i det følgende.

HVIS der er uafhængighed mellem rækker og søjler, må den relative fordeling af observationerne være den samme i hver række og i hver søjle:

*Eksempel:*
*Hvis tabellens* ***første søjle*** *indeholder i alt 10% af observationer i hele datamaterialet, må det – hvis der er uafhængighed mellem rækker og søjler – være sådan at...*

- *10% af observationer i første række, findes i første rækkes* ***første søjle***
- *10% af observationer i anden række, findes i anden rækkes* ***første søjle***
- *10% af observationer i r´te række, findes i r´te rækkes* ***første søjle***

*Og naturligvis tilsvarende for alle andre søjler i tabellen.*

## HVAD ER UAFHÆNGIGHED?



HVIS der er uafhængighed mellem tabellens rækker og søjler, og der derfor er den samme relative fordeling af observationerne i alle rækker og søjler, må vi forvente, at antallet af observationer i den $r$'te række og $c$'te søjle sådan cirka er

$$\begin{aligned} E_{r,c} &= \frac{\text{Antal observationer i } r\text{'te række}}{\text{Antal observationer i datamaterialet}} \\ &\quad \times \frac{\text{Antal observationer i } c\text{'te søjle}}{\text{Antal observationer i datamaterialet}} \\ &\quad \times \text{Antal observationer i datamaterialet} \\ &= \frac{\text{Antal observationer i } r\text{'te række} \times \text{Antal observationer i } c\text{'te søjle}}{\text{Antal observationer i datamaterialet}} \end{aligned}$$

## Hvad er uafhængighed?



HVIS der er uafhængighed mellem tabellens rækker og søjler, så bør...

- det forventede antal observationer $E_{r,c}$ i $r$'te række og $c$'te søjle
- og det faktiske antal observationer $X_{r,c}$ i $r$'te række og $c$'te søjle

være forholdsvis tæt på hinanden.

En tilstrækkelig stor afvigelse mellem $E_{r,c}$ og $X_{r,c}$, svarende til en tilstrækkelig stor værdi af $\left(E_{r,c} - X_{r,c}\right)^2$, vil således være i modstrid med, at der er uafhængighed mellem tabellens rækker og søjler.

Ved at se på (de kvadrerede) afvigelser i alle tabellens rækker og søjler får vi et mål for, om der ser ud til at være – eller ikke være – uafhængighed mellem rækker og søjler.

# Hvad er uafhængighed?



**Eksempel: Skat**

I spørgeskemaundersøgelsen om velfærd og skat kan vi undersøge, om der er en sammenhæng mellem, hvor i landet man er bosiddende, og om man synes, at topskatten er for høj.

Det gør vi ved at undersøge, om der kan antages at være uafhængighed mellem de to kategoriske variable *Region* og *"Er topskatten for høj?"*.

Inden vi laver et formelt test af hypotesen om uafhængighed, kan vi lave en hurtig og uformel sammenligning af datamaterialets observationer $X_{r,c}$ med de forventede størrelser $E_{r,c}$:

**Contingency Table**

| | Er topskatten for høj? | | |
|---|---|---|---|
| Count / Expected | Ja | Nej | Total |
| Region: Hovedstaden | 124 | 185 | 309 |
| | 109,655 | 199,345 | |
| Sjælland | 49 | 88 | 137 |
| | 48,6174 | 88,3826 | |
| Syddanmark | 66 | 136 | 202 |
| | 71,6841 | 130,316 | |
| Midtjylland | 80 | 149 | 229 |
| | 81,2656 | 147,734 | |
| Nordjylland | 27 | 71 | 98 |
| | 34,7774 | 63,2226 | |
| Total | 346 | 629 | 975 |

$X_{2,1}$ $E_{2,1}$ $X_{2,2}$ $E_{2,2}$

Af tabellen kan vi eksempelvis se, at blandt personerne bosiddende i Region Sjælland (række $r = 2$), ser det ud til, at de observerede og forventede tal (i tilfælde af at antagelsen om uafhængighed er korrekt) passer fint sammen. I alt $X_{2,1} = 49$ personer (række 2, søjle 1) mener, at topskatten er for høj, mens $X_{2,2} = 88$ ikke mener, at den er for høj.

Såfremt der ikke er sammenhæng mellem de to variable, så vil vi forvente, at $E_{2,1} = 48{,}6$ af de adspurgte personer i Region Sjælland mener, at topskatten er for høj og $E_{2,2} = 88{,}4$ mener, at den ikke er.

JMP-video [Analyze -> Fit Y by X]

## Resultat [Hypotesetest om uafhængighed: Trin 1-2]

**Trin 1:** *Antagelser*

- *To kategoriske variable med observationer i en $R \times C$ krydstabel med værdi $X_{r,c}$ i tabellens r´te række og c´te søjle*
- *Krydstabellen har den ene variabels $R$ værdier i rækkerne og den anden variabels $C$ værdier i søjlerne*
- *Alle variable $X_{r,c}$, $r = 1,\ldots,R$ $c = 1,\ldots,C$ er indbyrdes uafhængige*
- *$X_{r,c} \geq 5$ for alle rækker $r = 1,\ldots,R$ og alle søjler $c = 1,\ldots,C$*

**Trin 2:** *Hypotese*

$H_0$ : *Krydstabellens to variable er uafhængige*

$H_a$ : *Krydstabellens to variable er IKKE uafhængige*

Bemærk:
Vi betragter kun én bestemt nulhypotese $H_0$ her, nemlig nulhypotesen om uafhængighed. Der er ikke flere forskellige nulhypoteser at vælge imellem.

## Resultat [Hypotesetest om uafhængighed: Trin 3-5]

**Trin 3: Teststørrelse**

*Test af nulhypotesen $H_0$ udføres ved hjælp af teststørrelsen*

$$\mathbb{X} = \sum_{r=1}^{R} \sum_{c=1}^{C} \frac{(X_{r,c} - E_{r,c})^2}{E_{r,c}}$$

*Under forudsætning af at nulhypotesen $H_0$ er sand, er teststørrelsen $\mathbb{X}$ approksimativt beskrevet ved en $\chi^2$-fordeling med $f = (R-1)(C-1)$ frihedsgrader.*

**Trin 4: P-værdi**

*Store værdier af teststørrelsen $\mathbb{X}$ er i modstrid med nulhypotesen $H_0$, og P-værdien beregnes derfor som $P(\mathbb{X} > \mathbb{x})$, hvor $\mathbb{x}$ er værdien af teststørrelsen beregnet på baggrund af datamaterialet.*

**Trin 5: Konklusion**

*Hvis P-værdien er...*

- *mindre end signifikansniveauet $\alpha$ forkaster vi nulhypotesen $H_0$*
- *større end signifikansniveauet $\alpha$ forkaster vi ikke nulhypotesen $H_0$*

# HYPOTESETEST



**Eksempel: Skat**

Vi ser fortsat på de $n = 975$ indkomne svar på spørgsmålet *"Er topskatten for høj?"* opdelt på hvor i landet respondenterne er bosiddende (*"Region"*).

Hvis vi betragter hypoteserne

$H_0$ : Der er uafhængighed mellem bopælsregion og holdning til topskat
$H_a$ : Der er IKKE uafhængighed mellem bopælsregion og holdning til topskat

så svarer det til at undersøge, om der er sammenhæng mellem hvor i landet man er bosiddende og ens holdning til topskat.

Under forudsætning af at nulhypotesen $H_0$ er sand kan teststørrelsen beregnes til

$$\chi = 6{,}34$$

Tests

| N | DF | -LogLike | RSquare (U) |
|---|---|---|---|
| 975 | 4 | 3,2029401 | 0,0051 |

| Test | ChiSquare | Prob>ChiSq |
|---|---|---|
| Likelihood Ratio | 6,406 | 0,1708 |
| Pearson | 6,339 | 0,1752 |

$\chi$

P-værdi

Herefter kan P-værdien beregnes som sandsynligheden for at observere noget, der er i dårligere overensstemmelse med $H_0$ end værdien 6,34, dvs. som

$$P(\chi > 6{,}34) = 17{,}52\%$$

Ved et signifikansniveau på $\alpha = 5\%$ ***forkaster*** vi således ***ikke*** nulhypotesen $H_0$ (fordi P-værdi $= 17{,}52\% > 5\% = \alpha$). Der er således ***ikke*** på baggrund af datamaterialet belæg for at ***afvise***, at holdningen til topskat er den samme i hele landet.

JMP-video [Analyze -> Fit Y by X]

## $\chi^2$-FORDELINGEN



$\chi^2$-fordelingen ($\chi^2$ udtales "ki-i-anden") er vigtig, fordi den beskriver fordelingen af teststørrelsen, når vi laver test for uafhængighed mellem to kategoriske variable.

Nogle få facts om $\chi^2$-fordelingen. Fordelingen...

- har kun positive værdier (dvs. er altid $\geq 0$)
- har én parameter $f$, der kaldes "antal frihedsgrader"
- har centrum (middelværdi) $f$
- har standardafvigelse $\sqrt{2f}$
- minder om normalfordelingen, når $f$ bliver tilstrækkeligt stor
- bevæger sig til højre ud af 1. aksen, når $f$ vokser

Sandsynligheder og fraktiler i $\chi^2$-fordelingen kan beregnes i JMP på samme måde som for de øvrige fordelinger.

▶ JMP-video [Help -> Sample Data -> Teaching Scripts -> Interactive Teaching Modules -> Distribution Calculator]

## $\chi^2$-FORDELINGEN

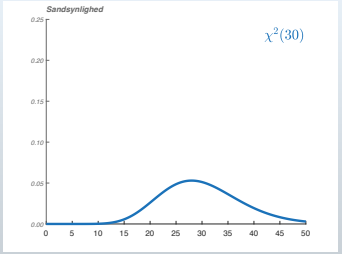


Nedenstående animation viser, hvordan $\chi^2$-fordelingen ændrer sig, i takt med at antallet af frihedsgrader $f$ øges.

Jo højere $f$ desto...

- længere ligger fordelingens centrum mod højre (= større middelværdi)
- bredere er fordelingen (= større standardafvigelse)

ANIMATION: $\chi^2$-fordelingen – ændring af antal frihedsgrader $f$




Start animation

Når vi i denne note taler om "hypotesetest om et antal middelværdier" betragter vi et datamateriale med observationer af ***én kategorisk variabel*** og ***én kvantitativ variabel***.

Den kategoriske variabel har $G$ mulige værdier, der bruges til at gruppere værdierne af den kvantitative variabel med. Tilfældet $G = 2$ svarer til analyse af to middelværdier $\mu_1$ og $\mu_2$ i note 7b.

Til forskel fra analysen i note 7b vil vi her udelukkende se på, om der ser ud til at være en sammenhæng mellem den kategoriske og den kvantitative variabel eller der ikke er. I tilfældet $G = 2$ svarer det til at se på hypoteserne

$$H_0 : \mu_1 - \mu_2 = 0 \qquad H_a : \mu_1 - \mu_2 \neq 0$$

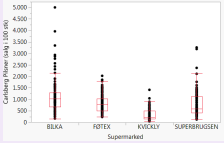
som vi også kan formulere som

$$H_0 : \mu_1 = \mu_2 \qquad H_a : \mu_1 \neq \mu_2$$

**Eksempel: Ølsalg**

Når vi ser på den ugentlige omsætning af Carlsberg (1 stk. 33 cl glasflaske) i de fire supermarkedskæder (= "grupper") Bilka, Føtex, Kvickly og SuperBrugsen, kan vi undersøge, om det forventede salg er det samme i alle fire kæder.

Inden vi laver et formelt test, kan vi lave en hurtig og uformel sammenligning af salget i de fire kæder ved at sammenligne box plots af fordelingen af salget indenfor hver kæde:




De optegnede box plots viser i store træk samme mønster, som vi allerede tidligere har set af tidsseriegrafer af omsætningen i hver af de fire kæder:

- Omsætningen i Kvickly ligger lavere end i de øvrige tre kæder
- Der er visse forskelle på variationen i den ugentlige omsætning fra kæde til kæde

JMP-video [Analyze -> Fit Y by X]

## Resultat [Hypotesetest om middelværdier $\mu_1, \ldots, \mu_G$: Trin 1-2]

**TRIN 1:** *Antagelser*

- *n observationer fordelt på G grupper med $n_g$ observationer i den g´te gruppe, $g = 1, \ldots, G$.*
- *Observationerne i den g´te gruppe $X_{g,1}, \ldots, X_{g,n_g}$ er approksimativt normalfordelt $N(\mu_g, \sigma)$, $g = 1, \ldots, G$*
- *alle datamaterialets i alt n observationer $(X_{1,j})_{j=1,\ldots,n_1}, \ldots, (X_{G,j})_{j=1,\ldots,n_G}$ er indbyrdes uafhængige*

**TRIN 2:** *Hypotese*

$H_0$ : $\mu_1 = \mu_2 = \ldots = \mu_G$

*(Nulhypotese: "Samme middelværdi i alle G grupper")*

$H_a$ : *Mindst to af middelværdierne $\mu_1, \ldots, \mu_G$ er ikke ens*

*(Alternativhypotese: "Ikke samme middelværdi i alle G grupper")*

BEMÆRK:

En af aftagelserne ovenfor er, at der er den samme variation (standardafvigelse $\sigma$) i alle G grupper.

## Resultat [Hypotesetest om middelværdier $\mu_1, \ldots, \mu_G$: Trin 3-5]

TRIN 3: *Teststørrelse*

*Test af nulhypotesen $H_0$ udføres ved hjælp af teststørrelsen*

$$\mathbb{F} = \frac{\text{“Variation i data indenfor hver af de } G \text{ grupper”}}{\text{“Variation mellem de } G \text{ grupper”}}$$

*Under forudsætning af at nulhypotesen $H_0$ er sand, er teststørrelsen $\mathbb{F}$ approksimativt beskrevet ved en $F$-fordeling med $(G-1, N-G)$ frihedsgrader.*

TRIN 4: *P-værdi*

*Store værdier af teststørrelsen $\mathbb{F}$ er i modstrid med nulhypotesen $H_0$, og P-værdien beregnes derfor som $P(\mathbb{F} > \mathbb{f})$, hvor $\mathbb{f}$ er værdien af teststørrelsen beregnet på baggrund af datamaterialet.*

TRIN 5: *Konklusion*

*Hvis P-værdien er...*

- *mindre end signifikansniveauet $\alpha$ forkaster vi nulhypotesen $H_0$*
- *større end signifikansniveauet $\alpha$ forkaster vi ikke nulhypotesen $H_0$*

BEMÆRK:

Vi vil ikke her gå nærmere ind i, hvordan teststørrelsen $\mathbb{F}$ beregnes og ej heller definitionen af teststørrelsens fordeling, den såkaldte $F$-fordeling.

**Eksempel: Ølsalg**

Vi ser fortsat på den ugentlige omsætning af Carlsberg i de fire forskellige supermarkedskæder Bilka, Føtex, Kvickly og SuperBrugsen.

Hvis vi betragter hypoteserne

$H_0 : \mu_{\text{Bilka}} = \mu_{\text{Føtex}} = \mu_{\text{Kvickly}} = \mu_{\text{SuperBrugsen}}$
$H_a$ : Mindst to af middelværdierne $\mu_{\text{Bilka}}, \mu_{\text{Føtex}}, \mu_{\text{Kvickly}}, \mu_{\text{SuperBrugsen}}$ er ikke ens

så svarer det til at undersøge, om den forventede ugentlige omsætning er den samme i alle fire kæder.

Under forudsætning af at nulhypotesen $H_0$ er sand kan teststørrelsen beregnes til

$$\mathbb{F} = 72{,}04$$

Herefter kan P-værdien beregnes som sandsynligheden for at observere noget, der er i dårligere overensstemmelse med $H_0$ end værdien 72,04, dvs. som

$$P(\mathbb{F} > 72{,}04) = \text{Mindre end } 0{,}01\%$$

**Oneway Anova**

**Summary of Fit**

| | |
|---|---|
| Rsquare | 0,257253 |
| Adj Rsquare | 0,253682 |
| Root Mean Square Error | 474,8075 |
| Mean of Response | 754,9929 |
| Observations (or Sum Wgts) | 628 |

**Analysis of Variance**

| Source | DF | Sum of Squares | Mean Square | F Ratio | Prob > F |
|---|---|---|---|---|---|
| Supermarked | 3 | 48723473 | 16241158 | 72,0414 | <,0001* |
| Error | 624 | 140675904 | 225442,15 | | |
| C. Total | 627 | 189399377 | | | |

$\mathbb{F}$ P-værdi

Ved et signifikansniveau på $\alpha = 5\%$ ***forkaster*** vi således nulhypotesen $H_0$ (fordi P-værdi $< 5\% = \alpha$). Der er således ***ikke*** på baggrund af datamaterialet belæg for at ***hævde***, at den forventede omsætning af Carlsberg er den samme i alle fire supermarkedskæder.

JMP-video [Analyze -> Fit Y by X]

1 Hypotesetest om uafhængighed

2 Hypotesetest om et antal middelværdier

# 3 OPSUMMERING

## *Kort opsummering af dette notesæt:*

Hypotesetest om et antal middelværdier $\mu_1, \ldots, \mu_G$

**Trin 1**: *Antagelser*

- $n$ observationer fordelt på $G$ grupper med $n_g$ observationer i den $g$'te gruppe, $g = 1, \ldots, G$.
- Observationerne i den $g$'te gruppe $X_{g,1}, \ldots, X_{g,n_g}$ er approksimativt normalfordelt $N(\mu_g, \sigma)$, $g = 1, \ldots, G$
- alle datamaterialets i alt $n$ observationer $(X_{1,j})_{j=1,\ldots,n_1}, \ldots, (X_{G,j})_{j=1,\ldots,n_G}$ er indbyrdes uafhængige

**Trin 2**: *Hypoteser*

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_G$
$H_a$ : Mindst to af middelværdierne $\mu_1, \ldots, \mu_G$ er ikke ens

**Trin 3**: *Teststørrelse*

Teststørrelsen

$$\mathbb{F} = \frac{\text{“Variation i data indenfor hver af de } G \text{ grupper”}}{\text{“Variation mellem de } G \text{ grupper”}}$$

er approksimativt beskrevet ved en $F$-fordeling med $(G-1, N-G)$ frihedsgrader, under forudsætning af at nulhypotesen $H_0$ er sand.

**Trin 4**: *P-værdi*

P-værdien beregnes som $P(\mathbb{F} > \mathbb{f})$, hvor $\mathbb{f}$ er værdien af teststørrelsen beregnet på baggrund af datamaterialet.

**Trin 5**: *Konklusion*

Hvis P-værdien er mindre (større) end signifikansniveauet $\alpha$ forkaster vi (forkaster vi ikke) nulhypotesen $H_0$.

## Hypotesetest om uafhængighed

**Trin 1**: *Antagelser*

- To kategoriske variable med observationer i en $R \times C$ krydstabel med værdi $X_{r,c}$ i tabellens $r$'te række og $c$'te søjle
- Krydstabellen har den ene variabels $R$ værdier i rækkerne og den anden variabels $C$ værdier i søjlerne
- Alle variable $X_{r,c}$, $\quad r = 1,\ldots,R$ $c = 1,\ldots,C$ er indbyrdes uafhængige
- $X_{r,c} \geq 5$ for alle rækker $r = 1,\ldots,R$ og alle søjler $c = 1,\ldots,C$

**Trin 2**: *Hypoteser*

$H_0$ : Krydstabellens to variable er uafhængige

$H_a$ : Krydstabellens to variable er IKKE uafhængige

**Trin 3**: *Teststørrelse*

Teststørrelsen

$$\mathbb{X} = \sum_{r=1}^{R} \sum_{c=1}^{C} \frac{(X_{r,c} - E_{r,c})^2}{E_{r,c}}$$

er approksimativt beskrevet ved en $\chi^2$-fordeling med $f = (R-1)(C-1)$ frihedsgrader, under forudsætning af at nulhypotesen $H_0$ er sand.

**Trin 4**: *P-værdi*

P-værdien beregnes som $P(\mathbb{X} > \mathrm{x})$, hvor $\mathrm{x}$ er værdien af teststørrelsen beregnet på baggrund af datamaterialet.

**Trin 5**: *Konklusion*

Hvis P-værdien er mindre (større) end signifikansniveauet $\alpha$ forkaster vi (forkaster vi ikke) nulhypotesen $H_0$.

# INDEKS

## Nye funktionaliteter i dette notesæt:

- *Analyze -> Fit Y by X:*
  - Test af hypotese om uafhængighed
  - Test af hypotese om $\mu_1, \ldots, \mu_G$

## JMP-videoer:

s. 2: ▶ [Graph -> Graph Builder] (grafik af uafhængighed)

s. 3: ▶ [Graph -> Graph Builder] (grafik af $\mu_1, \ldots, \mu_G$)

s. 7: ▶ [Analyze -> Fit Y by X] (test af hypotese om uafhængighed)

s. 15: ▶ [Help -> Sample Data -> Teaching Scripts -> Interactive Teaching Modules -> Distribution Calculator] (beregning i $\chi^2$-fordeling)

s. 18: ▶ [Analyze -> Fit Y by X] (test af hypotese om $\mu_1, \ldots, \mu_G$)