# HD Dataanalyse

*Notesæt 3: Sandsynlighedsberegning*

*(i krydstabel)*

Copenhagen Business School

# Emne i dette notesæt



Når vi har et datamateriale med to kategoriske variable og ønsker at undersøge sammenhængen mellem dem, er det mest naturlige at optælle, hvor mange observationer vi har af hver kombination af de to variable. Denne optælling opskriver vi i en tabel og kalder tabellen for en krydstabel, fordi den "krydser´ ´ (dvs. kombinerer) de to variable.

På baggrund af tabellen vil vi gerne kunne udtale os om, hvor sandsynligt det er at forskellige kombinationer af de to variable optræder i vores datamateriale. Med andre ord så vil vi gerne kunne formulere udsagn af formen: *"Vi kan se af tabellen, at der er …% sandsynlighed for at …"*.

Det er imidlertid ikke umiddelbart klart, hvordan man helt præcist skal formulere udsagn af formen "sandsynligheden for at …" på baggrund af en krydstabel. Notesæt 3 beskæftiger sig derfor med, hvorledes man på korrekt vis beregner og fortolker sandsynligheder på baggrund af en krydstabel.

Bemærk: De ting, vi ser på her i notesæt 3, gælder mere generelt end blot for hyppigheder i krydstabeller. Vi vil imidlertid her i notesæt 3 udelukkende anvende krydstabeller til at eksemplificere de forskellige regneregler for sandsynligheder med, fordi det er den mest intuitive måde at forstå tingene på.

# Emne i dette notesæt



**Eksempel: Boligpriser**

På baggrund af de kategoriske variable *Boligtype* og *Landsdel* kan vi opstille en krydstabel med hyppigheder. Vi ser her for nemheds skyld kun på boliger i de tre landsdele København (= Kbh. & Frederiksberg), Vestjylland og Østjylland.

| | | Boligtype | | | | |
|---|---|---|---|---|---|---|
| | Count | Ejerlejlighed | Fritidshus | Rækkehus | Villa | Total |
| Landsdel | Kbh & Frederiksberg | 10839 | 14 | 2374 | 7230 | 20457 |
| | Vestjylland | 374 | 526 | 302 | 4782 | 5984 |
| | Østjylland | 2767 | 815 | 1019 | 7950 | 12551 |
| | Total | 13980 | 1355 | 3695 | 19962 | 38992 |

Spørgsmålet er nu, hvad vi kan konkludere på baggrund af denne tabel.

Hvis vi tilfældigt udvælger én bolig fra datamaterialet, hvad er så eksempelvis...

- sandsynligheden for at boligen er en villa?
- sandsynligheden for at boligen er en villa og ligger i København?
- sandsynligheden for at boligen ligger i København, hvis vi ved at boligen er et rækkehus?
- sandsynligheden for at boligen ligger i København, hvis vi ved at boligen ikke er et rækkehus?

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

# Indholdsfortegnelse

## BEREGNING AF SANDSYNLIGHEDER

I forbindelse med beregning af sandsynligheder anvender man følgende grundlæggende begreber:

- **eksperiment** betegner det, vi ønsker at undersøge
  *(f.eks. hvad man får i karakter ved næste eksamen)*
- **udfaldsrum** *("sample space")* betegner alle de mulige udfald vores eksperiment kan resultere i
  *(karakteren ved næste eksamen har udfaldsrummet* $\{-3, 00, 02, 4, 7, 10, 12\}$*)*
- **hændelse** *("event")* betegner en samling af et eller flere mulige resultater af vores eksperiment
  *(hændelsen "at bestå eksamen" svarer f.eks. til* $\{02, 4, 7, 10, 12\}$*, mens hændelsen "at bestå med højeste eller næsthøjeste karakter" svarer til* $\{10, 12\}$*)*

Bemærk:

- Enhver hændelse er en del af udfaldsrummet (fordi udfaldsrummet består af *alle* de mulige resultater af vores eksperiment).
- Hændelser er de ting, vi ønsker at beregne sandsynligheder for.
- En hændelse betegnes som regel notationsmæssigt med et stort bogstav ($A$, $B, C, \ldots$).
- Sandsynligheden for en hændelse $A$ betegnes med $P(A)$.
  *(hvis $A$ betegner hændelsen "at bestå eksamen", så er $A = \{02, 4, 7, 10, 12\}$ og sandsynligheden for at bestå eksamen betegnes da $P(A)$)*

**Eksempel: Boligpriser**

Et eksempel på et eksperiment kan være, at vi tilfældigt udvælger én ud af de solgte boliger i vores datamateriale og så ser på, hvilken type boligen er.

Eksperimentets udfaldsrum er da {ejerlejlighed, villa, fritidshus, rækkehus}, mens eksempler på hændelser kunne være...

- $A =$"boligen er ikke en lejlighed", eller alternativt formuleret $A =$" boligen er en villa, et fritidshus eller et rækkehus"
- $A =$"boligen hænger bygningsmæssigt sammen med andre boliger", eller alternativt formuleret $A =$"boligen er en ejerlejlighed eller et rækkehus"

Beregning af sandsynligheder på baggrund af et datamateriale foregår på følgende måde:

**Definition (Beregning af sandsynlighed)**

*På baggrund af et datamateriale beregnes* ***sandsynligheden*** *("probability") for en hændelse $A$ som*

$$P(A) = \frac{\text{antal observationer af hændelsen } A \text{ i datamaterialet}}{\text{antal observationer i datamaterialet}}$$

Bemærk:

- Den ovenstående metode til beregning af sandsynligheder sikrer, at der altid vil gælde at $0 \leq P(A) \leq 1$, uanset hvilken hændelse $A$ vi ser på.
- Udfaldsrummet for et eksperiment har per definition altid sandsynlighed 1, dvs. $P(\text{udfaldsrum}) = 1$

  (fordi eksperimentet altid vil få et resultat, som er en del af udfaldsrummet)

# BEREGNING AF SANDSYNLIGHEDER

**Eksempel: Boligpriser**

Krydstabellen for de kategoriske variable *Boligtype* og *Landsdel* har følgende form:

| Landsdel | Boligtype: Count | Ejerlejlighed | Fritidshus | Rækkehus | Villa | Total |
|---|---|---|---|---|---|---|
| | Kbh & Frederiksberg | 10839 | 14 | 2374 | 7230 | 20457 |
| | Vestjylland | 374 | 526 | 302 | 4782 | 5984 |
| | Østjylland | 2767 | 815 | 1019 | 7950 | 12551 |
| | Total | 13980 | 1355 | 3695 | 19962 | 38992 |

Hvis vi betragter hændelsen $A =$"boligen er en villa", så kan vi beregne sandsynligheden

$$P(A) = \frac{\text{antallet af villaer i datamaterialet}}{\text{antallet af observationer i datamaterialet}} = \frac{19.962}{38.992} = 51,20\%$$

Tallet 51,20% udtrykker sandsynligheden for, at en vilkårligt valgt bolig blandt samtlige solgte boliger i de tre landsdele København, Vestjylland og Østjylland er en villa. Alternativt formuleret, at andelen af villaer blandt samtlige solgte boliger i de tre landsdele er 51,20%.

Hvis vi i stedet betragter hændelsen $B =$"boligen ligger i København", så kan vi beregne sandsynligheden

$$P(B) = \frac{\text{antallet af boliger i København i datamaterialet}}{\text{antallet af observationer i datamaterialet}} = \frac{20.457}{38.992} = 52,46\%$$

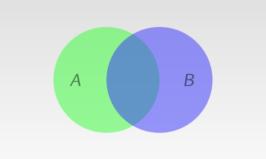
Tallet 52,46% udtrykker sandsynligheden for, at en vilkårligt valgt bolig blandt samtlige solgte boliger i de tre landsdele er beliggende i København. Alternativt formuleret, at andelen af samtlige solgte boliger i de tre landsdele, som er solgt i København er 52,46%.

Når man skal forstå de forskellige regneregler, der gælder for sandsynligheder, er det ofte lettest at tegne tingene op i en figur, et såkaldt **Venn diagram**.

Ideen er, at den grafiske fremstilling i Venn diagrammet skal gøre det lettere at forstå, hvorfor de enkelte regneregler ser ud, som de gør.

Ofte er vi interesseret i at beregne sandsynligheder, der involverer en kombination af to (eller flere) hændelser. Et Venn diagram, der involverer to hændelser $A$ og $B$ tegnes på følgende måde:

- først tegnes udfaldsrummet
- dernæst tegnes hændelsen $A$
- dernæst tegnes hændelsen $B$




Venn diagrammet giver en grafisk illustration af de forskellige ting, vi kan beregne sandsynligheder for, på baggrund af hændelserne $A$ og $B$.

Beregning
Sandsynlighed
Betinget
OPSUMMERING

Regning med sandsynligheder har sin egen, lidt specielle notation i form af de tre symboler $\cap$, $\cup$, $^c$, som har hver sin sproglige betydning.

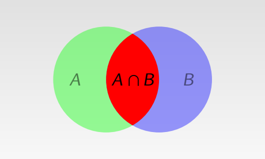
Ved først øjekast kan brugen af symbolerne godt virke en smule besværlig og unødig indviklet.

Årsagen til at man bruger symbolerne er, at det viser sig at være en matematisk set "smart" måde at formulere sig på, når man skal beregne sandsynligheder for komplicerede hændelser, der i sig selv er kombinationer af andre hændelser.

For at kunne formulere de forskellige regneregler for sandsynligheder er vi nødt til først at forstå, hvad de tre symboler $\cap$, $\cup$, $^c$ betyder.

## Fælleshændelsen *("intersection")* $A \cap B$:

- Symbolet $\cap$ har den sproglige betydning "***og***"
- Fælleshændelsen $A \cap B$ læses derfor som "$A$ og $B$"
- Fælleshændelsen $A \cap B$ er den del af udfaldsrummet, som dækkes af både $A$ og $B$




**Eksempel: Boligpriser**

Hvis vi betragter hændelserne

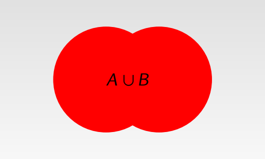
$$A = \text{"boligen er en villa"} \qquad B = \text{"boligen er beliggende i København"}$$

så er fælleshændelsen

$$A \cap B = \text{"boligen er en villa } \textbf{\textit{og}} \text{ den er beliggende i København"}$$

## Foreningshændelsen *("union")* $A \cup B$:

- Symbolet $\cup$ har den sproglige betydning "***eller***"
- Foreningshændelsen $A \cup B$ læses derfor som "$A$ eller $B$ (eller både $A$ og $B$)"
- Foreningshændelsen $A \cup B$ er den del af udfaldsrummet, som dækkes af enten $A$ eller $B$ eller af både $A$ og $B$




**Eksempel: Boligpriser**

Hvis vi betragter hændelserne

$$A = \text{"boligen er en villa"} \qquad B = \text{"boligen er beliggende i København"}$$

så er foreningshændelsen

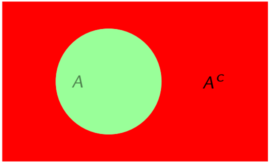
$$A \cup B = \text{"boligen er enten en villa } \textbf{\textit{eller}} \text{ den er beliggende i København"}$$

Foreningshændelsen betegner således samtlige villaer (uanset landsdel) samt alle øvrige boligtyper (dvs. ejerlejligheder, fritidshuse og rækkehuse) i det omfang de er beliggende i København.

Beregning
Sandsynlighed
Betinget
OPSUMMERING

## Komplementærhændelsen *("complement")* $A^c$:

- Symbolet $^c$ har den sproglige betydning "***ikke***"
- Komplementærhændelsen $A^C$ læses derfor som "er ikke $A$"
- Komplementærhændelsen $A^C$ er den del af udfaldsrummet, som ikke dækkes af $A$




**Eksempel: Boligpriser**

Hvis vi betragter hændelserne

$$A = \text{"boligen er en villa"} \qquad B = \text{"boligen er beliggende i København"}$$

så er komplementærhændelsen til $A$

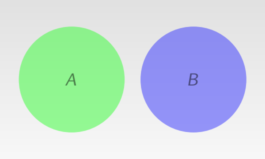
$$\begin{aligned} A^c &= \text{"boligen er ikke en villa"} \\ &= \text{"boligen er en ejerlejlighed, et fritidshus eller et rækkehus"} \end{aligned}$$

Tilsvarende er komplementærhændelsen til $B$

$$\begin{aligned} B^c &= \text{"boligen er ikke beliggende i København"} \\ &= \text{"boligen er beliggende i Vestjylland eller Østjylland"} \end{aligned}$$

## Disjunkte *("disjoint")* hændelser:

- To hændelser $A$ og $B$ kaldes disjunkte, hvis de ikke har noget til fælles




**Eksempel: Boligpriser**

Hvis vi betragter hændelserne

$$A = \text{"boligen er en villa"} \qquad B = \text{"boligen er en ejerlejlighed"}$$

så er hændelserne disjunkte, fordi en bolig ikke både kan være en villa og en ejerlejlighed.

Tilsvarende hvis vi betragter hændelserne

$$A = \text{"boligen er en villa i København"} \qquad B = \text{"boligen er en villa i Østjylland"}$$

så er hændelserne ligeledes disjunkte, fordi en villa ikke på samme tid kan ligge både i København og i Østjylland.

Beregning
Sandsynlighed
Betinget
OPSUMMERING

I forbindelse med brugen af symbolerne $\cap$, $\cup$, $^c$ kan det ind i mellem være nyttigt at kende til nedenstående sammenhænge.

## Resultat [Sammenhænge mellem hændelser]

*For to hændelser $A$ og $B$ gælder følgende:*

- $A \cap B = B \cap A$

  *(Intuition: det der "både er i $A$ og $B$" er det samme som det, der "både er i $B$ og $A$")*

- $A \cup B = B \cup A$

  *(Intuition: det der "enten er i $A$ eller $B$" er det samme som det, der "enten er i $B$ eller $A$")*

- $((A^c)^c) = A$

  *(Intuition: det der "ikke er det, der ikke er i $A$" er det samme som $A$)*

- $(A \cap B)^c = A^c \cup B^c$

  *(Intuition: det "der ikke både er i $A$ og $B$" er det samme som det, der "enten ikke er i $A$ eller ikke er i $B$")*

- $(A \cup B)^c = A^c \cap B^c$

  *(Intuition: det "der ikke enten er i $A$ eller $B$" er det samme som det, der "både ikke er i $A$ og ikke er i $B$")*

**Eksempel: Boligpriser**

Hvis vi betragter hændelserne

$$A = \text{“boligen er en villa”} \qquad B = \text{“boligen er beliggende i København”}$$

så er eksempelvis de to fælleshændelser $A \cap B$ og $B \cap A$ givet som

$$A \cap B = \text{“boligen er en villa og boligen er beliggende i København”}$$
$$B \cap A = \text{“boligen er beliggende i København og boligen er en villa”}$$

og det er klart, at de to hændelser beskriver præcis de samme boliger, nemlig villaer i København.

Ligeledes er eksempelvis

$$(A^c)^c = \text{“boligen er ikke en bolig, der ikke er beliggende i København”}$$

hvilket blot er en (noget besværlig) beskrivelse af netop de boliger, der er beliggende i København, dvs. $(A^c)^c = A$.

Vi er nu endelig klar til at formulere de grundlæggende regneregler for sandsynligheder.

**Resultat [Regneregler for sandsynligheder]**

*For to vilkårlige hændelser $A$ og $B$ gælder at...*

- $P(A^c) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

*For to **disjunkte** hændelser $A$ og $B$ gælder at...*

- $P(A \cap B) = 0$
- $P(A \cup B) = P(A) + P(B)$

BEMÆRK: Regnereglerne ovenfor viser, hvordan man kan finde sandsynlighederne $P(A^c)$ og $P(A \cup B)$ ud fra kendskab til sandsynlighederne $P(A), P(B), P(A \cap B)$.

**Eksempel: Boligpriser**

Lad os igen betragte krydstabellen for de kategoriske variable *Boligtype* og *Landsdel*.

| Landsdel / Count | Ejerlejlighed | Fritidshus | Rækkehus | Villa | Total |
|---|---|---|---|---|---|
| Kbh & Frederiksberg | 10839 | 14 | 2374 | 7230 | 20457 |
| Vestjylland | 374 | 526 | 302 | 4782 | 5984 |
| Østjylland | 2767 | 815 | 1019 | 7950 | 12551 |
| Total | 13980 | 1355 | 3695 | 19962 | 38992 |

Hvis vi betragter hændelserne

$$A = \text{“boligen er en villa”} \qquad B = \text{“boligen er beliggende i København”}$$

har vi tidligere set at $P(A) = 51,20\%$ og $P(B) = 52,46\%$.

Vi kan dermed beregne sandsynligheden for, at en tilfældigt valgt bolig ikke er en villa som

$$P(A^c) = 1 - P(A) = 1 - 51,20\% = 48,80\%$$

Bemærk, at denne sandsynlighed alternativt også kan beregnes direkte ud fra krydstabellen som

$$P(A^c) = \frac{13.980 + 1.355 + 3.695}{38.992} = 48,80\%$$

**Eksempel: Boligpriser (fortsat)**

På baggrund af krydstabellen

| Landsdel / Count | Ejerlejlighed | Fritidshus | Rækkehus | Villa | Total |
|---|---|---|---|---|---|
| Kbh & Frederiksberg | 10839 | 14 | 2374 | 7230 | 20457 |
| Vestjylland | 374 | 526 | 302 | 4782 | 5984 |
| Østjylland | 2767 | 815 | 1019 | 7950 | 12551 |
| Total | 13980 | 1355 | 3695 | 19962 | 38992 |

kan vi f.eks. også beregne sandsynligheden for, at en tilfældigt valgt bolig er en villa beliggende i København som

$$P(A \cap B) = \frac{7.230}{38.992} = 18,54\%$$

og dermed sandsynligheden for at en tilfældigt valgt bolig er en villa eller er beliggende i København som

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 51,20\% + 52,46\% - 18,54\% = 85,12\%$$

Denne sandsynlighed kan også alternativt beregnes direkte ud fra tabellen som

$$P(A \cup B) = \frac{10.839 + 14 + 2.374 + 7.230 + 4.782 + 7.950}{38.992} = 85,12\%$$

Beregning
Sandsynlighed
Betinget
OPSUMMERING

Nogle gange har vi behov for at udtrykke/beregne sandsynligheder, der er baseret på noget ***ekstra information***.

Sådanne sandsynligheder kaldes for ***betingede sandsynligheder***, fordi det er sandsynligheder der udregnes, hvor vi ***betinger*** med noget ekstra information.

Når man beregner betingede sandsynligheder, svarer det til, at man gør udfaldsrummet mindre, fordi den information man betinger med, udelukker visse dele af det oprindelige udfaldsrum.

Betingede sandsynligheder optræder ofte i formuleringer som

*"sandsynligheden for ... **givet** [en eller anden betingelse]"*

eller

*"sandsynligheden for ... **hvis** [en eller anden betingelse]"*

Ordene "givet" og "hvis" refererer til den ekstra information, der betinges på og er det, der afslører, at der er tale om en betinget sandsynlighed og ikke bare en almindelig sandsynlighed.

# Betingede sandsynligheder

**Eksempel: Boligpriser**

Hvis vi eksempelvis vil beregne *sandsynligheden for, at en tilfældigt valgt bolig er beliggende i København,* så ser vi på antallet af boliger beliggende i København i forhold til det samlede antal boliger i datamaterialet:

$$\frac{\text{antallet af } \textbf{\textit{samtlige boliger}} \text{ beliggende i København}}{\text{antallet af } \textbf{\textit{samtlige boliger}} \text{ i datamaterialet}}$$

Hvis vi stedet vil beregne *sandsynligheden for, at en tilfældigt valgt bolig er beliggende i København, givet at boligen er et rækkehus,* dvs. vi forudsætter her (betinger på) at boligen er et rækkehus, så ser vi i stedet på antallet af rækkehuse beliggende i København i forhold til det samlede antal rækkehuse i datamaterialet:

$$\frac{\text{antallet af } \textbf{\textit{rækkehuse}} \text{ beliggende i København}}{\text{antallet af } \textbf{\textit{rækkehuse}} \text{ i datamaterialet}}$$

Sidstnævnte er et eksempel på en *betinget sandsynlighed,* fordi vi betinger sandsynligheden på den ekstra information, at boligerne vi ser på skal være rækkehuse.

**Definition (Betinget sandsynlighed)**

*For to hændelser $A$ og $B$ beregnes den **betingede sandsynlighed** ("“conditional probability”) **for** $A$ **givet** $B$ som*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Den betingede sandsynlighed $P(A|B)$ angiver sandsynligheden for, at hændelsen $A$ indtræffer, ***givet*** at vi allerede ved, at hændelsen $B$ er indtruffet.

Bemærk: Der gælder altid at...

- $0 \le P(A|B) \le 1$
- $P(A|A) = 1$
- $P(A|B) = 0$ hvis $A$ og $B$ er disjunkte

# BETINGEDE SANDSYNLIGHEDER

**Eksempel: Boligpriser**

Lad os fortsat betragte krydstabellen for de kategoriske variable *Boligtype* og *Landsdel*.

| Landsdel \ Boligtype: Count | Ejerlejlighed | Fritidshus | Rækkehus | Villa | Total |
|---|---|---|---|---|---|
| Kbh & Frederiksberg | 10839 | 14 | 2374 | 7230 | 20457 |
| Vestjylland | 374 | 526 | 302 | 4782 | 5984 |
| Østjylland | 2767 | 815 | 1019 | 7950 | 12551 |
| Total | 13980 | 1355 | 3695 | 19962 | 38992 |

Hvis vi betragter hændelserne

$$A = \text{“boligen er et rækkehus”} \qquad B = \text{“boligen er beliggende i København”}$$

så er $P(A) = \frac{3.695}{38.992} = 9,48\%$ og $P(B) = \frac{20.457}{38.992} = 52,46\%$ og $P(A \cap B) = \frac{2.374}{38.992} = 6,09\%$.

Vi kan dermed beregne sandsynligheden for, at et tilfældigt valgt rækkehus er beliggende i København som

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{6,09\%}{52,46\%} = 11,61\%$$

Det vil sige, at hvis vi kun ser på solgte boliger i København, så er 11,61% $(= P(A|B))$ af dem rækkehuse. Ser vi derimod på samtlige solgte boliger i de tre landsdele, så er 9,48% $(= P(A))$ af dem rækkehuse.

Andelen af rækkehuse blandt de solgte boliger er således lidt større i København end den i gennemsnit er på tværs af de tre landsdele. M.a.o. så tyder det på, at der alt andet lige sælges relativt flere rækkehuse i København end i Vest- og Østjylland (når vi betragter Vest- og Østjylland under ét).

**Eksempel: Boligpriser (fortsat)**

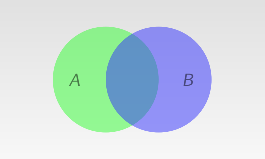
Bemærk også forskellen på de to sandsynligheder $P(A|B)$ og $P(A \cap B)$.

Den betingede sandsynlighed $P(A|B) = 11{,}61\%$ vedrører udelukkende boliger solgt i København (= hændelsen $B$), mens den almindelige sandsynlighed $P(A \cap B) = 6{,}09\%$ vedrører samtlige solgte boliger i de tre landsdele (= hele udfaldsrummet).

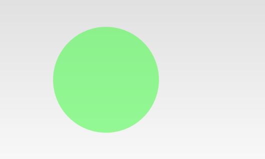
Blandt boliger solgt i København er 11,61% rækkehuse, mens blandt samtlige boliger solgt i de tre landsdele er 6,09% rækkehuse beliggende i København.

# BETINGEDE SANDSYNLIGHEDER

For bedre at forstå forskellen på den betingede sandsynlighed $P(A|B)$ og den almindelige sandsynlighed $P(A)$ kan vi se på Venn diagrammet fra tidligere.




$\underline{P(A)}$:
Ved beregningen af den almindelige sandsynlighed $P(A)$ betragter vi hele det oprindelige udfaldsrum og ser så på, hvor stor en del hændelsen $A$ udgør af det.



$\underline{P(A|B)}$:
Ved beregningen af den betingede sandsynlighed $P(A|B)$ begrænser vi udfaldsrummet til at være hændelsen $B$ og ser så på, hvor stor en del hændelsen $A$ udgør af det.

**Resultat [Regneregler for betingede sandsynligheder]**

*For to vilkårlige hændelser $A$ og $B$ gælder at...*

- $P(A \cap B) = P(B) \cdot P(A|B)$
- $P(A \cap B) = P(A) \cdot P(B|A)$

**Eksempel: Boligpriser**

Hvis vi betragter hændelserne

$$A = \text{“boligen er et rækkehus”} \qquad B = \text{“boligen er beliggende i København”}$$

så har vi set at

$$P(B) = 52{,}46\%, \qquad P(A|B) = 11{,}61\% \qquad P(A \cap B) = 6{,}09\%$$

hvilket netop passer med regnereglen ovenfor, idet

$$6{,}09\% = P(A \cap B) = P(B) \cdot P(A|B) = 52{,}46\% \cdot 11{,}61\%$$

**Definition (Uafhængighed)**

*To hændelser $A$ og $B$ kaldes* ***uafhængige*** *("independent"), hvis hændelserne ikke påvirker hinandens sandsynligheder.*

Et eksempel på hændelser, der er uafhængige, er i forbindelse med kast med en terning.

Hvis en terning kastes to gange, og vi betragter hændelserne

$$A = \text{“6'er i 1. terningkast”} \qquad B = \text{“6'er i 2. terningkast”}$$

så er de uafhængige, fordi hvorvidt man får en 6′er i første terningkast eller ej (hændelse $A$) påvirker ikke, hvorvidt man får en 6′er i andet terningkast (hændelse $B$).

# BETINGEDE SANDSYNLIGHEDER

**Eksempel: Boligpriser**

Intuitivt betyder uafhængighed mellem to hændelser $A$ og $B$, at den ene hændelse ikke indeholder information om den anden hændelse og vice versa.

Eksempelvis vil hændelserne

$$A = \text{"boligen er et rækkehus"} \qquad B = \text{"boligen er beliggende i København"}$$

være uafhængige, hvis det er sådan, at andelen af rækkehuse er den samme for boliger hhv. i og udenfor København. I det tilfælde vil sandsynligheden for at en solgt bolig er et rækkehus nemlig være den samme, uanset om boligen er beliggende i København eller ej.

Hvis der derimod er forskel på andelen af rækkehuse blandt solgte boliger hhv. i og udenfor København, så ligger der noget information om sandsynligheden for at en solgt bolig er et rækkehus (tilhører $A$) i at vide, hvorvidt boligen ligger i København eller ej (tilhører $B$ eller ej).

### Resultat [Kriterier for uafhængighed]

*De to hændelser $A$ og $B$ er uafhængige, hvis blot én af følgende betingelser er opfyldt:*

- $P(A \cap B) = P(A) \cdot P(B)$
- $P(A|B) = P(A)$
- $P(B|A) = P(B)$

**Eksempel: Boligpriser**

For hændelserne

$$A = \text{“boligen er et rækkehus”} \qquad B = \text{“boligen er beliggende i København”}$$

har vi set at

$$9,48\% = P(A) \neq P(A|B) = 11,61\%$$

og dermed er hændelserne $A$ og $B$ ikke uafhængige.

Intuitivt er forklaringen den samme, som vi tidligere har set på, nemlig at rækkehuse udgør en relativt større del af den samlede mængde solgte boliger i København, end tilfældet er i de øvrige landsdele, vi betragter (Vest- og Østjylland).

Når vi vil undersøge, hvor sandsynligt det er, at en tilfældigt valgt solgt bolig er et rækkehus og samtidig betinger på, at den solgte bolig skal være beliggende i København (dvs. når vi beregner $P(A|B)$), så er det alt andet lige mere sandsynligt, at boligen vil være et rækkehus, end hvis vi ser på solgte boliger i alle de tre betragtede landsdele (dvs. når vi beregner $P(A)$).

# OPSUMMERING

## *Kort opsummering af dette notesæt:*

Sandsynligheden for en hændelse $A$ beregnes som $P(A) = \frac{\text{antal observationer af } A}{\text{samlet antal observationer}}$.

De vigtigste typer af hændelser er:

- Fælleshændelse $(A \cap B)$
- Foreningshændelse $(A \cup B)$
- Komplementærhændelse $(A^c)$

To hændelser $A$ og $B$ er disjunkte, hvis de ikke har noget til fælles, og i så fald er $P(A \cap B) = 0$.

For regning med sandsynligheder gælder at...

- $P(A^c) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

En betinget sandsynlighed $P(A|B) = \frac{P(A \cap B)}{P(B)}$ er en sandsynlighed, hvor der betinges med informationen i hændelsen $B$.

For betingede sandsynligheder gælder at $P(A \cap B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$.

To hændelser $A$ og $B$ er uafhængige, hvis én af følgende betingelser er opfyldt:

- $P(A \cap B) = P(A) \cdot P(B)$
- $P(A|B) = P(A)$
- $P(B|A) = P(B)$

# INDEKS

## Nye funktionaliteter i dette notesæt:

Ingen

## JMP-videoer:

s. 2: ▶ [Analyze -> Fit Y by X] (krydstabel)