Sandsynlighedsregning 3. forelæsning

Bo Friis Nielsen

Anvendt Matematik og Computer Science

Danmarks Tekniske Universitet

2800 Kgs. Lyngby – Danmark

Email: bfni@dtu.dk

Dagens emner

- Stokastiske variable: *udfald identificeres med reelle tal*
  - Fordeling $\mathsf{P}(X = x)$ $\quad \sum_x \mathsf{P}(X = x) = 1$
  - Simultan(joint) fordeling
    $P(x, y) = \mathsf{P}(X = x, Y = y) \quad \sum_{x,y} \mathsf{P}(X = x, Y = y) = 1$
  - Marginal fordeling
    $P_X(x) = \mathsf{P}(X = x) = \sum_y \mathsf{P}(X = x, Y = y)$
  - Betinget fordeling $\mathsf{P}(Y = y | X = x) = \frac{\mathsf{P}_{(X=x,Y=y)}}{\mathsf{P}_{(X=x)}} = \frac{P(x,y)}{P_X(x)}$
  - Uafhængighed $P(x, y) = P_X(x) P_Y(y)$
    $\mathsf{P}(Y = y, X = x) = \mathsf{P}(X = x)\mathsf{P}(Y = y), \ \forall (x, y)$
- Multinomialfordelingen
- Middelværdi $\mathsf{E}(X) = \sum x \cdot \mathsf{P}(X = x)$
- Forventningsværdi mere generelt $\mathsf{E}(g(X)) = \sum g(x) \cdot \mathsf{P}(X = x)$
- Linearitet $\mathsf{E}(aX + bY + c) = a\mathsf{E}(X) + b\mathsf{E}(Y) + c$

Stokastiske variable kap.3

- Beskriver eksperimenter, hvor udfaldet kan karakteriseres ved et eller flere tal
- Gør det muligt at definere gennemsnit og andre beregninger
- Gør notation og formulering mere smidig

Eksempel på anvendelse

- For binomialfordelingen indførte vi hændelserne S_i for i succeser
- Hændelsen H_i højst i successer blev skrevet som $\cup_{j=0}^{i} S_j$
- Så $\mathsf{P}(H_i) = \mathsf{P}\left(\cup_{j=0}^{i} S_j\right) = \sum_{j=0}^{i} \mathsf{P}(S_j)$ Hvorfor?
- Vi kunne ønske mere kompakt notation, færre symboler
- Den stokastiske variabel X betegner antallet af succeser
- Hvordan vil vi skrive hændelsen S_5 ved hjælp af X?
- Hvordan vil vi skrive hændelsen H_5 ved hjælp af X?
- En stokastisk variabel er en "funktion", der beskriver udfaldet af vores eksperiment, før vi har udført det.

Fordele ved stokastiske variable

- Mange problemer formuleres mere enkelt
- Giver os mulighed for at danne nye stokastiske variable ved funktioner
- Knytter en forbindelse mellem udfaldsrummet og de reelle tal
- Formelt $\mathsf{P}\left(X = x\right) = \mathsf{P}\left(\{\omega \in \Omega : X(\omega) = x\}\right)$
  Originalmængden/urbilledet
- $\mathsf{P}\left(X \leq x\right) = \mathsf{P}\left(\{\omega \in \Omega : X(\omega) \leq x\}\right)$

Vi lader den stokastiske variabel X betegne antallet af krone i tre kast med en retfærdig mønt, og definerer en ny stokastisk variabel $Y = |X - 1|$.

Spørgsmål 1

Angiv sandsynligheden for $\mathsf{P}(Y = 1)$.

1 ☐ $\frac{1}{8}$

2 ☐ $\frac{1}{4}$

3 ☐ $\frac{3}{8}$

4 ☐ $\frac{1}{2}$

5 ☐ $\frac{5}{8}$

6 ☐ Ved ikke

Københavnske boliger 1989 (antal)

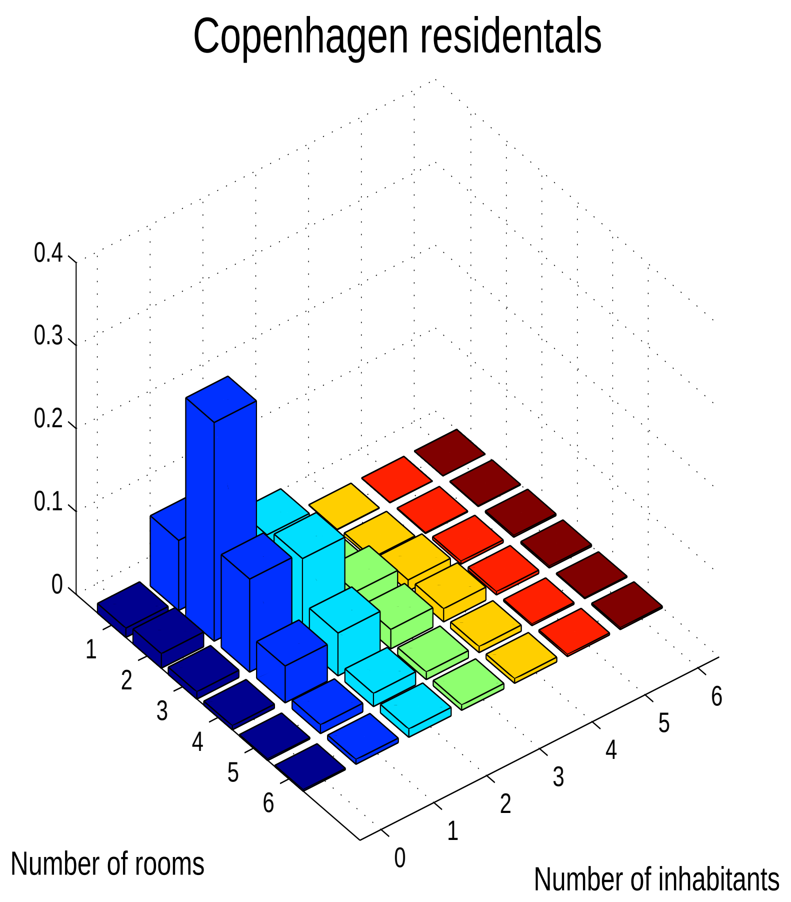
Antal beboere:	0	1	2	3	4	5	6
1 værelse	3166	23554	2038	260	69	19	8
2 værelser	4874	73508	26670	3826	955	248	145
3 værelser	2408	31193	29023	8781	3735	925	565
4 værelser	1343	12301	14374	6487	4380	1064	585
5 værelser	436	2877	4376	2487	1977	538	264
6 værelser	442	1553	2748	1775	1543	615	521

- En bolig er således her karakteriseret ved to egenskaber størrelse og antal beboere

Relative andele

Antal beboere	0	1	2	3	4	5	6
1 værelse	0,011	0,085	0,007	0,001	0,000	0,000	0,000
2 værelser	0,017	0,264	0,096	0,014	0,003	0,001	0,001
3 værelser	0,009	0,112	0,104	0,032	0,013	0,003	0,002
4 værelser	0,005	0,044	0,052	0,023	0,016	0,004	0,002
5 værelser	0,002	0,010	0,016	0,009	0,007	0,002	0,001
6 værelser	0,002	0,006	0,010	0,006	0,006	0,002	0,002

Histogram over bolighyppigheder



Flerdimensionale stokastiske variable

- Vi indfører nu symbolet X til at betegne antallet af værelser i en tilfældigt valgt bolig og Y til at betegne antallet af beboere i denne bolig.
- En bolig er således karakteriseret ved parret (X, Y) - en todimensional stokastisk variabel.
- Vi har $\mathsf{P}(X = x) = P_X(x)$ til at beskrive antallet af værelser i en bolig
- Og $\mathsf{P}(Y = y) = P_Y(y)$ til at beskrive antallet af beboere i en bolig.
- Som en ny ting får vi brug for $P(x, y) = \mathsf{P}(X = x, Y = y)$ til at beskrive de samlede egenskaber ved en bolig.
- Den simultane fordeling

Brug af den simultane fordeling

- Hvad er sandsynligheden for, at en tilfældigt valgt bolig har højst to værelser og højst 3 beboere?

$$\sum_{x=1}^{2}\sum_{y=0}^{3} P(x,y) = 0,011 + \ldots 0,014 = 0,4978$$

- Hvad er sandsynligheden for, at en tilfældigt valgt bolig vil have flere værelser end beboere?

$$\sum_{x=1}^{6}\sum_{y=0}^{x-1} P(x,y) = 0,011+0,017+0,264+0,009+\cdots+0,002 = 0,7160$$

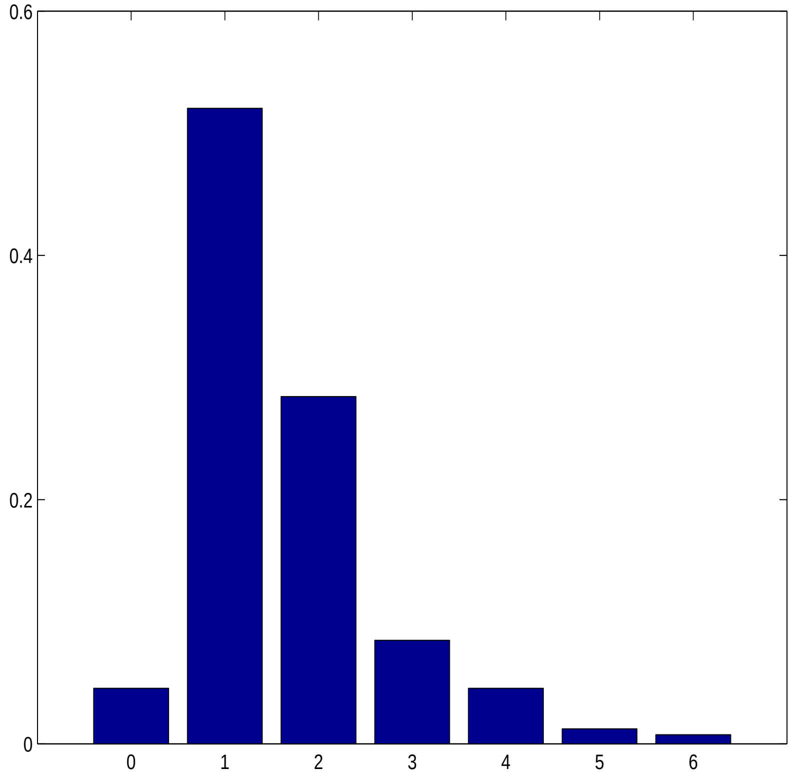
Hvis vi kun interesserer os for antal beboere

Antal beboere	0	1	2	3	4	5	6
1 værelse	0,011	0,085	0,007	0,001	0,000	0,000	0,000
2 værelser	0,017	0,264	0,096	0,014	0,003	0,001	0,001
3 værelser	0,009	0,112	0,104	0,032	0,013	0,003	0,002
4 værelser	0,005	0,044	0,052	0,023	0,016	0,004	0,002
5 værelser	0,002	0,010	0,016	0,009	0,007	0,002	0,001
6 værelser	0,002	0,006	0,010	0,006	0,006	0,002	0,002
Boliger i alt	0,045	0,520	0,284	0,085	0,045	0,012	0,007

- Dvs. kun Y

- altså Y

$$\mathsf{P}(Y=y)=\sum_x \mathsf{P}(X=x,Y=y)=\sum_x P(x,y)=P_Y(y)$$

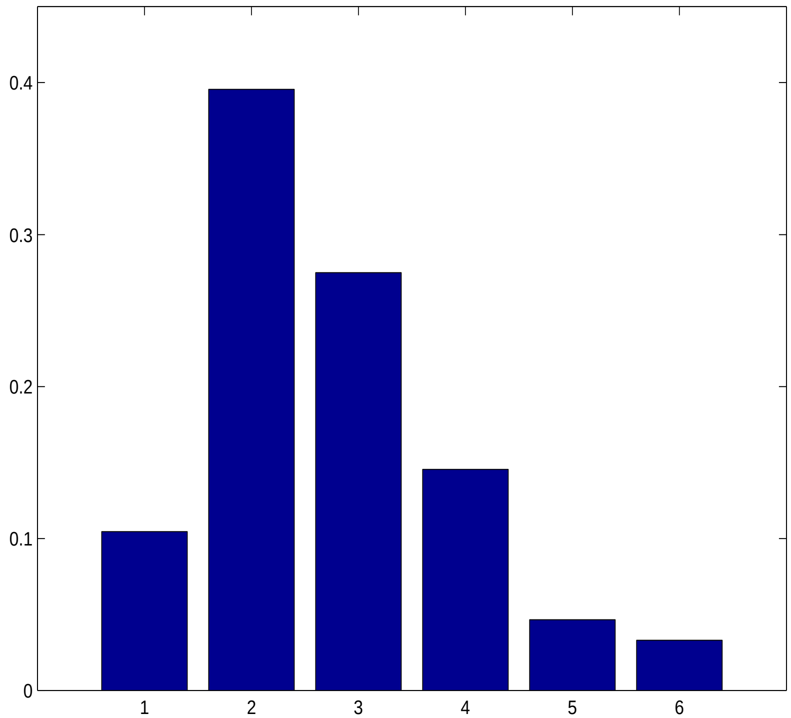


- Den marginale fordeling af Y $(P_Y(y))$

Hvis vi kun interesserer os for størrelsen

- I så fald interesserer vi os for X

$$\mathsf{P}(X=x)=\sum_y \mathsf{P}(X=x,Y=y)=\sum_y P(x,y)=P_X(x)$$



- Den marginale fordeling af X $(P_X(x))$

Antallet af beboere i en 3 værelses bolig

- Hvis vi ved, at en bolig har 3 værelser, betyder det så noget for fordelingen af an-
  tal beboere? Fra rækken med 3 værelses boliger i tabellerne har vi:

Beboere:	0	1	2	3	4	5	6	Alle
Antal:	2408	31193	29023	8781	3735	925	565	76630
Andel:	0,009	0,112	0,104	0,032	0,013	0,003	0,002	0,275
Andel af 3 værelses	0,031	0,407	0,379	0,115	0,049	0,012	0,007	1,000

- Hvordan udtrykker vi dette symbolsk/matematisk?

Betinget sandsynlighed/fordeling,

For hændelser A og B har vi definitionen

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

Så for hændelserne $X = x$ og $Y = y$ får vi

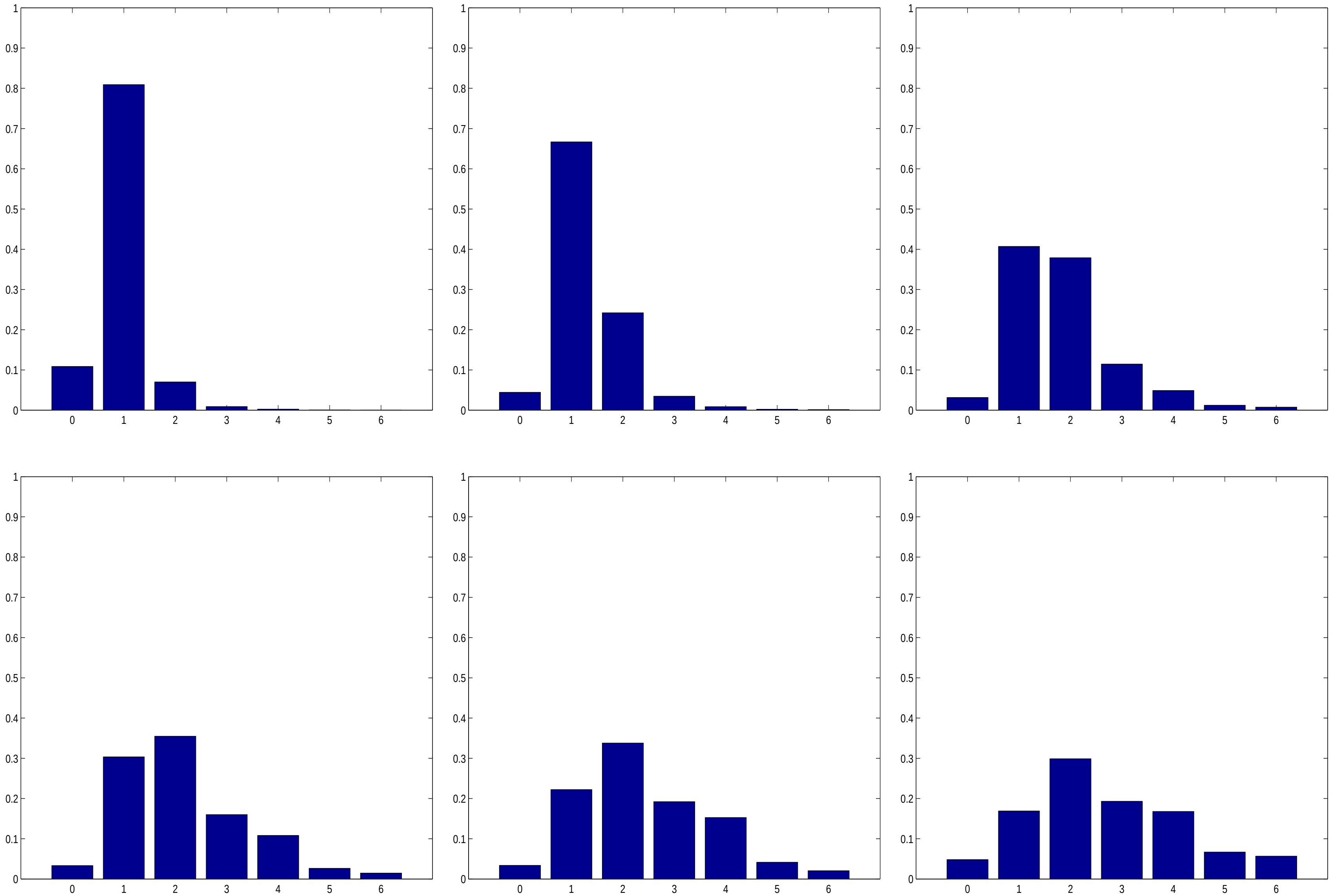
$$\mathsf{P}(Y = y|X = x) = \frac{\mathsf{P}(X = x, Y = y)}{\mathsf{P}(X = x)}$$

- Den betingede fordeling. Konkret får vi

$$\mathsf{P}(Y = y|X = 3) = \frac{\mathsf{P}(X = 3, Y = y)}{\mathsf{P}(X = 3)} = \frac{P(3, y)}{0,275}$$

som i tabellen

Beboere betinget af størrelse



Uafhængige stokastiske variable

- Vi har set, hvordan den betingede fordeling fremkommer ved hjælp af betingede sandsynligheder.
- Nogle gange ved vi intuitivt (antagelse), at en størrelse (stokastisk variabel) ikke vekselvirker med en anden stokastisk variabel.
  - Terningeeksempler
- Hvad er en naturlig definition på uafhængige stokastiske variable?
- Er X og Y uafhængige i boligeksemplet?

Definition på uafhængige stokastiske variable

- X og Y er uafhængige hvis

$$\mathsf{P}(X = x, Y = y) = \mathsf{P}(X = x)\mathsf{P}(Y = y)$$

eller udtrykt med P-funktionerne

$$P(x, y) = P_X(x)P_Y(y)$$

- for alle x, y
- Giv et eksempel på uafhængige X og Y (gode kandidater)

Opgave 3.12

- Lad X og Y være resultatet (talværdi) fra to trækninger fra en æske med 4 sedler, med numrene 1,2,3 og 4.
- Angiv den marginale fordeling og simultane fordeling for X og Y. For
  - ◇ Stikprøvetagning med tilbagelægning
  - ◇ Stikprøvetagning uden tilbagelægning
- Beregn P$(X \leq Y)$
- Er X og Y uafhængige?

Med tilbagelægning får vi

$X \backslash Y$	1	2	3	4
1	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
2	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$

$$\mathsf{P}(X = i) = \mathsf{P}(Y = i) = \frac{1}{4} = P_X(i) = P_Y(i)$$

$$\mathsf{P}(X \leq Y) = \sum_{x \leq y} P(x, y) = \sum_{x=1}^{4} \sum_{y=x}^{4} P(x, y)$$

$$= \sum_{y=1}^{4} \sum_{x=1}^{y} P(x, y) = \sum_{y=1}^{4} y \frac{1}{16} = \frac{10}{16} = \frac{5}{8}$$

Uden tilbagelægning får vi

$$\mathsf{P}(X = i) = \frac{1}{4} = P_X(i)$$

$X \backslash Y$	1	2	3	4
1	0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$
2	$\frac{1}{12}$	0	$\frac{1}{12}$	$\frac{1}{12}$
3	$\frac{1}{12}$	$\frac{1}{12}$	0	$\frac{1}{12}$
4	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	0

$$\mathsf{P}(Y = i) = \sum_{x=1}^{4} P(x, i) = \frac{1}{4} = P_Y(i)$$

Uden tilbagelægning fortsat

$$\mathsf{P}(X \leq Y) = \sum_{x \leq y} P(x,y) = \sum_{x=1}^{4} \sum_{y=x}^{4} P(x,y)$$

$$= \sum_{y=1}^{4} \sum_{x=1}^{y} P(x,y) \sum_{y=1}^{4} \sum_{x=1}^{y-1} P(x,y)$$

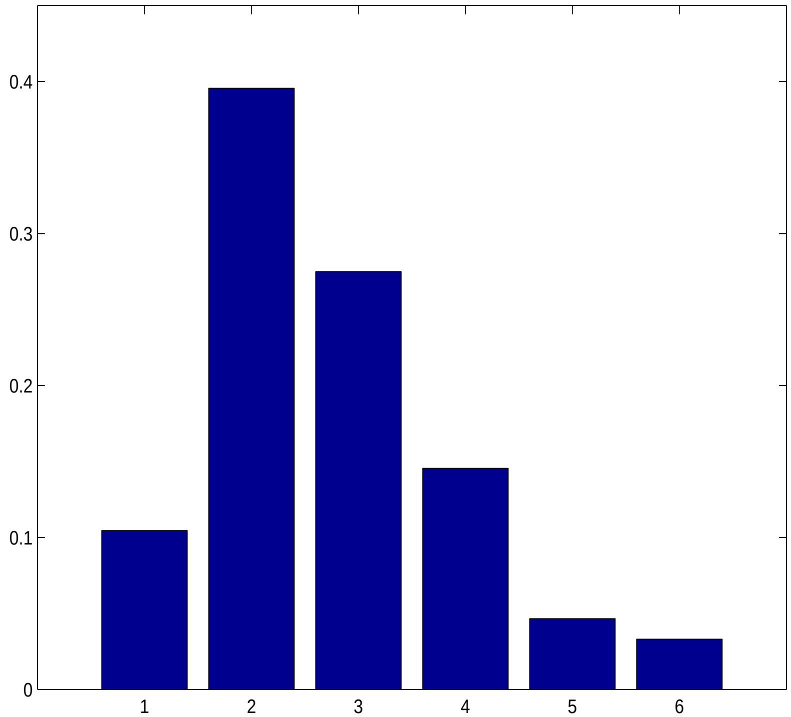
$$= \sum_{y=1}^{4} (y-1) \frac{1}{12} = \frac{6}{12} = \frac{1}{2}$$

- Kunne vi have indset dette direkte?

Funktioner af stokastiske variable

- Fra boligeksemplet:
  - Z_1 : antal beboere pr. værelse $Z_1 = \frac{Y}{X}$
  - Z_2 : antal værelser pr. beboer $Z_2 = \frac{X}{Y} = \frac{1}{Z_1}$ (uproblematisk?)

Fordeling af værelser på antal boliger



værelser:	1	2	3	4	5	≥ 6
andel:	0,104	0,396	0,275	0,145	0,046	0,033

Enkel karakteristik af fordeling

Spørgsmål 2

Hvilket af følgende tal ville du vælge til at karakterisere fordelingen af antal værelser i en bolig?

1 ☐ 2

2 ☐ 3

3 ☐ 3.5

4 ☐ $1 \cdot 0,104 + 2 \cdot 0,396 + 3 \cdot 0,275 + 4 \cdot 0,145 + 5 \cdot 0,046 + 6 \cdot 0,033$
$= 2,729$

5 ☐ Noget andet

6 ☐ Ved ikke

Middelværdi(forventning) p.162(163)

$$\mathsf{E}(X) = \sum_{\text{alle x}} x\mathsf{P}(X = x)$$

- Fordelingens tyngdepunkt
- Gennemsnit af stort antal udfald, eksperimenter
- Beregningsmæssig bekvem

Spørgsmål 3

Hvad er middelværdien/forventningsværdien ved kast med en terning?

1 ☐ Alle værdier $\{1, 2, 3, 4, 5, 6\}$

2 ☐ 3

3 ☐ 3,5

4 ☐ 4

5 ☐ 4,5

6 ☐ Ved ikke

Bemærk

- Middelværdien af X, $\mathsf{E}(X)$ behøver ikke at være en mulig værdi for X

Skatteindtægt fra bolig

- Skatteindtægten fra en bolig afhænger af antallet af værelser kr. 2.500 pr. værelse.
- Plus et fast beløb - kr. 7.500
- Hvad er den forventede skatteindtægt fra en bolig?

Lineære funktioner af stokastiske variable

$$\mathsf{E}(aX + b) = a\mathsf{E}(X) + b$$

- i Eksemplet får vi: Z : skatteindtægt fra bolig

$$\mathsf{E}(Z) = \mathsf{E}(2.500 \cdot X + 7.500)$$

$$= 2.500 \cdot \mathsf{E}(X) + 7.500 = 14.322,50$$

Forventning af en funktion af en stokastisk variabel p.175

- Lidt mere generelt får vi for en vilkårlig funktion $g(x)$:

$$\mathsf{E}(g(X)) = \sum_{\text{alle } x} g(x)\mathsf{P}(X = x)$$

- Af særlig interesse har vi $g(x) = x^k$, så

$$\mathsf{E}(X^k) = \sum_{\text{alle } x} x^k\mathsf{P}(X = x)$$

- det k'te moment af X

Antag, at $\mathsf{E}(X^2) = 3$, $\mathsf{E}(Y^2) = 4$, $\mathsf{E}(XY) = 2$.

Spørgsmål 4

$E\left[(X+Y)^2\right]$ findes til

1 ☐ 7

2 ☐ 9

3 ☐ 11

4 ☐ 13

5 ☐ 15

6 ☐ Ved ikke

Forventning af en funktion af flere stokastiske variable p.176

- Generelt har vi:

$$\mathsf{E}(g(X,Y)) = \sum_{\text{alle } (x,y)} g(x,y)\mathsf{P}(X = x, Y = y)$$

- Additionsregel $\mathsf{E}(X+Y) = \mathsf{E}(X) + \mathsf{E}(Y)$ (altid) p.167.
- Multiplikationsregel, hvis X og Y er uafhængige p.177

$$\mathsf{E}(XY) = \mathsf{E}(X)\mathsf{E}(Y)$$

Relative andele af boliger

Antal beboere	0	1	2	3	4	5	6
1 værelse	0,011	0,085	0,007	0,001	0,000	0,000	0 ,000
2 værelser	0,017	0,264	0,096	0,014	0,003	0,001	0 ,001
3 værelser	0,009	0,112	0,104	0,032	0,013	0,003	0 ,002
4 værelser	0,005	0,044	0,052	0,023	0,016	0,004	0 ,002
5 værelser	0,002	0,010	0,016	0,009	0,007	0,002	0 ,001
6 værelser	0,002	0,006	0,010	0,006	0,006	0,002	0 ,002

- Hvordan ville I beregne den forventede værdi af antallet af beboere pr. værelse?

Sammensætning/funktioner af stokastiske variable

- Først definerer vi X som antal værelser i en bolig, dernæst Y som antallet af beboere.
- Vi danner en ny variabel $Z = \frac{Y}{X}$ som antal beboere pr. værelse

$$\mathsf{E}(Z) = \mathsf{E}(Y/X) = \sum_{x=1}^{6} \sum_{y=0}^{6} \frac{y}{x} P(x,y)$$

Nye begreber i afsnit 3.1, 3.2

- Stokastisk variabel
- Fordeling/marginal fordeling
- Flerdimensional stokastisk variabel
- Simultan (eng.: joint) fordeling
- Betinget fordeling
- Uafhængige stokastiske variable
- Middelværdi/Forventning

Afsnit 3.1 og 3.2

- Stokastiske variable: *udfald identificeres med relle tal*
  - Fordeling $\mathsf{P}(X=x) \qquad \sum_x \mathsf{P}(X=x)=1$
  - Simultan(joint) fordeling
    $P(x,y)=\mathsf{P}(X=x,Y=y) \qquad \sum_{x,y}\mathsf{P}(X=x,Y=y)=1$
  - Marginal fordeling
    $P_X(x)=\mathsf{P}(X=x)=\sum_y \mathsf{P}(X=x,Y=y)$
  - Betinget fordeling $\mathsf{P}(Y=y|X=x)=\frac{\mathsf{P}(X=x,Y=y)}{\mathsf{P}(X=x)}=\frac{P(x,y)}{P_X(x)}$
  - Uafhængighed $\mathsf{P}(Y=y,X=x)=\mathsf{P}(X=x)\mathsf{P}(Y=y)$
- Multinomialfordelingen
- Middelværdi $\mathsf{E}(X)=\sum x\cdot\mathsf{P}(X=x)$
- Forventningsværdi mere generelt $\mathsf{E}(g(X))=\sum g(x)\cdot\mathsf{P}(X=x)$
- Linearitet $\mathsf{E}(aX+bY+c)=a\mathsf{E}(X)+b\mathsf{E}(Y)+c$